

7

Noise and Variability in Experimental Data

7.1 “Noise” in Economics and in Experimental Economics

While much of this book has been concerned with broad issues of methodological principle, this chapter addresses what might seem to be a more specific and practical question associated with the analysis and interpretation of experimental data.

Proponents of experimental economics often claim that experiments are particularly well suited to testing economic theory, since they give a much greater degree of control than can generally be found in the “natural” economic environment. While that may be true, this chapter will suggest that insufficient consideration has so far been given to the question of which tests are suitable for different kinds of experimental data and under what conditions particular null hypotheses are appropriate. There seems to be a widespread assumption that statistical and econometric techniques developed for use with nonexperimental data can be applied quite straightforwardly to the data generated by experiments. But is this a safe assumption? This chapter will suggest that it is not, and will show that this apparently practical problem of statistical analysis connects back to some of the broader methodological issues raised earlier in the book.

Consider first the general approach underlying much econometric work. Economic theory has traditionally developed models where one variable—say the quantity demanded of good X —is a function of other variables (such as the price of X , consumers’ incomes, the prices of complements and substitutes for X , and so on). As presented by theorists in standard mathematical forms, these are for the most part deterministic models: that is to say, the model is written as if the dependent variable is completely and precisely determined by the values of the independent variables. Econometricians then take whatever data are available about those different variables in order to examine how well particular models

work and to obtain numerical estimates of the coefficients that express the impacts of the different independent variables upon the dependent variable.

When undertaking this task, they typically face a number of problems. First, the ways in which variables interact may not be fully specified by the model—or if they are, the specification may not correspond exactly with the ways in which those variables actually do interact. Second, there may be many variables that actually have *some* effect on the value of the dependent variable, but many of these effects may be relatively small and/or there may be little or no available data about some of them. Third, the data for the most significant variables may be imperfect: they may be liable to measurement error, and, if they are derived from some sampling procedure, they will also be subject to sampling error. Fourth (although this may be regarded by some as a subset of the second category), there may be heterogeneity in the preferences and motivations of different actors in the economy that is difficult to observe directly. Fifth, there may be some “within-subject” variability, so that even when presented with what might appear to be exactly the same conditions, agents may act differently on different occasions. All of these may be sources of “noise” and “error” that make it difficult to identify and quantify the underlying core relationships with total precision.

Many econometric studies therefore focus on a manageable set of what theory suggests are the most important independent variables and then model all other influences as if their collective impact upon the dependent variable operates much like random error. Within this framework, tests may be conducted to see whether particular variables are significantly influential and how different models, or different specifications of those models, compare. As part of this process, the robustness of the assumptions about the error term may also be investigated, and tests (or the conclusions drawn from them) may be modified in the light of the results of those investigations.

However, there are differences between experimental and nonexperimental data that may mean, at the very least, that careful thought should be given to the ways in which statistical and econometric techniques developed for nonexperimental data are applied to experimental results. In particular, the fifth source of noise listed above—within-subject variability in expressed or revealed preferences—does not feature very prominently as a separate consideration in most econometric applications; and even the fourth source—variability of tastes/goals *between* agents—may be swamped by the first three categories. But precisely because experiments are designed to control many of the other sources of noise in nonexperimental data, variability within and between

actors' beliefs, preferences, and judgments may play a larger and more central role in the generation of experimental data; and this may have significant implications for the analysis and interpretation of those data.

In the nonexperimental world of decision making under risk, people operate on the basis of their own perceptions and evaluations of consequences and their own subjective estimates of the risks, which may vary widely in ways that are hard to observe. By contrast, individual decision experiments typically involve a small number of clearly specified payoffs and use probabilities that correspond with some explicit random mechanism, as explained in chapter 2. In the nonexperimental world of strategic interaction, requirements of game theory (such as common knowledge) are rarely satisfied and it is hard to say what different players know or believe about the strategies and payoffs available to other players. By contrast, in the laboratory, experimenters often go to considerable lengths to try to make it common knowledge among the players what the strategies available to each player are and how payoffs depend upon them, at least in monetary terms.¹ In nonexperimental markets, it is often the case that little is known about the cost functions of suppliers or the true willingness to buy of purchasers across a wide range of prices; and even when some knowledge of these things can be obtained, it is rare to find two or more market institutions operating under comparable demand and supply conditions. By contrast, market experiments may induce the same underlying demand and supply functions and then study how different institutions perform under those conditions (see chapter 3 for details of inducing values in markets).

In short, in classic theory-testing experiments, the conventional experimental wisdom is to aim for as much control as possible over the experimental environment and then to introduce human actors into that environment and observe their behavior. Under such circumstances, the noise from other sources is greatly reduced: the principal—or at least, a major—element in any stochastic component in the data is likely to derive from the participants themselves.² Thus, modeling the stochastic component in the data, and deciding upon the appropriate tests to conduct, requires us to model the sources and nature of noise and error in human judgment and decision making. And it turns out that

¹Strictly speaking, the standard game-theoretic assumption is that payoffs are common knowledge *in utility form*. The methods by which monetary payoffs and other features of the game are made common knowledge—for example, publicly announcing a set of instructions in which these matters are described—cannot actually make payoffs in utility terms common knowledge. As will be discussed in section 7.3, not having control over *these* features is liable to be a source of "noise."

²Since any underlying population of economic actors is liable to be heterogeneous, sampling error is also likely to be an element.

different “error stories” can have quite different implications for the hypotheses we consider and for the validity of the statistical tests used to discriminate between them.

The next section will discuss three different forms of error story that have been proposed in the context of individual decision making under risk. Section 7.3 will consider some issues and implications for experiments examining game theory. The final section will reflect upon some possible strategies for adding to our understanding in this challenging and underdeveloped area of experimental methodology.

In what follows, we deliberately restrict attention to what might be thought of as individuals’ inherent uncertainty/imprecision/propensity to error when considering their own preferences and/or weighting of probabilities. There is also a substantial literature concerned with the ways in which people may learn about the experimental environment they are placed in and/or the behavior of others whose actions affect them—for example, how they may adapt to the strategies of other players in the course of a repeated experimental game. Camerer (2003, chapter 6) provides a very useful review of learning models, many of which take a stochastic form. Our focus, however, is upon variability in response that cannot very easily be attributed to feedback and the acquisition of new information, but that seems intrinsic to the human judgmental apparatus.

7.2 “Noise” in Individual Decision Experiments

The focus in this section will be upon individual decision making under risk. In fact, the scope of individual decision experiments is a good deal wider than that (see Camerer 1995) but decision making under risk is a convenient place to start because, as discussed in chapters 2–4, there is a wealth of material from studies investigating seemingly systematic departures from standard theory and/or trying to compare the performance of one decision theory against another.

One particularly pertinent feature of *some* of these studies is that they provide evidence about what happens when the same individuals are asked to undertake exactly the same decision task on two or more occasions within a short space of time—i.e., either within the same experimental session or in repeated sessions a day or two apart. The bulk of this evidence relates to pairwise choices. For example, Loomes and Sugden (1998) reported a study where, in the first part of each experimental session, they asked respondents to make a series of forty-five choices between pairs of simple lotteries involving no more than three payoffs.

They then gave the respondents a short "distractor" task before presenting them a few minutes later with the same forty-five pairwise choices again. The only difference between the two series was the order in which the questions were presented, which was randomized on both occasions; but all other features—not least the particular way in which each pair was set out—were kept exactly the same in both series, in order to minimize any noise due to presentational or "framing" effects. In order to control for income effects, the random lottery incentive system, discussed in chapter 6 and defined in box 6.1 (p. 265), was used.

Despite these efforts to keep the tasks as similar as possible across stages, there were many instances where an individual chose one alternative on the first occasion that they were asked to make a decision and chose the other alternative when they were presented with exactly the same pair of lotteries for a second time. In fact, out of a total of 3,680 comparisons (ninety-two respondents each making forty choices³ on two occasions), there were 676 cases (18.4%) where the choice on the second occasion was different from that on the first.⁴ To what extent might such behavior be regarded as white noise or random error? And to the extent that it is, how should it be understood and modeled?

One possibility is that what is being observed here is not error, but rather some kind of learning. Could it be that, in facing the decision tasks the first time round, the respondents gradually learned how to deal with them more rationally, or gradually discovered their "true" preferences, and that their choices the second time round revealed the benefits of this experience? (Such an interpretation would be consistent with the discovered preference hypothesis, discussed in chapter 2 and defined in box 2.7 (p. 76).)

In fact, there is *some* evidence that behavior changed systematically over the course of the experiment. In each of the forty pairs, one lottery could be regarded as the "safer" (*S*) option while the other was the "riskier" (*R*) option, with the latter usually offering a higher expected value than the former. If there were only purely random disturbances in respondents' choices in the first series and a corresponding pattern of random disturbance in the second series, we might expect as many cases where *S* was chosen in the first series and *R* in the second series

³We focus here on the forty choices that did not involve dominance and exclude the five choices where one option strictly dominated the other. More will be said about those choices later.

⁴These rates of reversal are by no means untypical: for some time it has been accepted that in similar circumstances, when neither lottery is obviously inferior to the other, as many as 30% of respondents may choose differently when presented with the same pair on two occasions a few minutes apart (see, for example, Camerer 1989; Starmer and Sugden 1989; Hey and Orme 1994).

(denoted by *SR*) as cases where *R* was chosen first time round and *S* on the second occasion (denoted by *RS*). But this was not so. From the total of 676 instances where *S* was chosen on one occasion and *R* on the other, 407 were *RS* as opposed to 269 *SR*. If the null hypothesis were that all reversals of choice between the first and second occasions were due *only* to random error such that *SR* was just as likely to be observed as *RS*, a standard binomial test would reject that null hypothesis with a very high degree of confidence. Rather, there appears to have been some trend toward making safer choices on the second occasion than the first—a trend that also turned out to be present in the data reported by Hey and Orme (1994) and by Ballinger and Wilcox (1997). However, even if the excess of *RS* over *SR* reversals is attributed to learning, more than 14% of the observations remain to be accounted for. This suggests that there is a considerable degree of purely stochastic variation in these data.⁵

To try to capture such variability, three sorts of stochastic model have been most prominent in experimental economics in recent years. We shall refer to these as the *Tremble model*, the *Fechner model*, and the *Random Preference (RP) model*. These are described in the following subsections, together with a discussion of their strengths and limitations, and some of their respective implications for the interpretation of different patterns of data from experiments examining individual decision making under risk.⁶

7.2.1 The Tremble Model

Perhaps the simplest notion of error is that people have some “true” preferences, but occasionally make mistakes in reporting them, due possibly to momentary inattention or some similar lapse—in short, due to a slip or a “trembling hand.” For pairwise choices, this could be formalized by supposing that an individual truly prefers one option but that there is some probability ω that she will make a slip in the course of translating preference into choice and will be observed to choose the truly less-preferred option. Intuitively, it seems implausible that ω is totally independent of the characteristics of the pair of alternatives being

⁵Loomes et al. (2002) use econometric methods to try to disentangle the effects of stochastic variation and learning in the data reported by Loomes and Sugden (1998). In detail, their conclusions depend on the way stochastic variation is modeled, but the broad picture is clear: there is a significant learning effect, but even at the end of the ninety decision tasks, individual responses show a high degree of stochastic variation.

⁶Our discussion draws on Loomes (2005) and reflects a line of research going back to Loomes and Sugden (1995). A different, and very useful, review of the approaches we consider here has recently been provided by Wilcox (2008).

considered. However, when this specification was used by Harless and Camerer (1994) to facilitate a meta-analysis of studies of violations of independence, they had insufficient detailed information about patterns of choice at the level of individual respondents, so that the rather crude assumption of the same ω being applied to all choices was convenient for processing the data in the aggregated form available to them.

A somewhat more refined approach was used by Sopher and Gigliotti (1993) when investigating the extent to which nontransitive choice cycles might be attributable to error. In their study they had data at the level of individual respondents and were able to allow ω to vary from one choice problem to another, and they found some evidence that there was such variability. Later, Loomes and Sugden (1998) used the data referred to in the previous subsection to test the very restrictive assumption that ω was the same across all problems and strongly rejected it.⁷

So while there may be circumstances where the data do not allow anything more sophisticated, the Tremble model has limited intuitive or empirical appeal. It is also difficult to see how it might be used for some other kinds of decision problems, such as those involving certainty equivalent valuations. Suppose it is truly the case that an individual's certainty equivalent for a particular lottery is \$5. What would/could the Tremble model say about the chances that other values will be reported? How *many* other values might be reported as a result of slips? And would each of those slips be equally likely to occur?

However, since there is already good evidence that this model performs poorly even in relation to pairwise choice tasks, there seems little point in investing time and effort into generalizing it to a broader set of tasks. That is not to say that slips and momentary inattention play *no part at all* in decision data. When respondents are being asked to make a large number of choices of a somewhat unfamiliar kind in a relatively short period of time, it would be surprising if there were no such trembles; and indeed, Loomes et al. (2002) argued that a "tremble term" is a useful adjunct to other ways of specifying the stochastic element in choice. But it is not viable as the principal model.

7.2.2 The Fechner Model

Gustav Fechner (1860) was one of the founding fathers of psychophysics. The particular focus of psychophysics is the way human subjects judge

⁷ The proposition was clearly rejected on the basis of the forty pairs that did not involve dominance. Had the tests also considered the five pairs where one lottery dominated the other, the result would have been even stronger: by contrast with the 18.4% average reversal rate across the forty "no dominance" pairs, the average rate for the five pairs involving dominance was just 2.4%.

the magnitudes of physical stimuli: in particular, how subjective judgments of magnitude map to “objective” measures.⁸ For example, is a sound that involves twice as much energy perceived by human subjects to be twice as loud? Is one weight that is actually twice as heavy as another perceived to be so?

One of the things that emerged very early in the course of such research was the imperfection of human judgment. Take the case of judging weight and imagine the following experiment. The respondent is blindfolded and seated at a table. Two objects are placed on the table, each with a cord attached. The respondent is told to lift each object in turn by the cord (so that any differences in shape or texture are neutralized) and then is asked to judge which of the two is heavier.

Suppose that one object weighs 500 grams and the other weighs 510 grams: how likely is it that the subject will correctly judge the 510 gram object to be heavier? The evidence is that when the difference between two stimuli is relatively small, there is a significant probability (though less than 0.5) of making a mistaken judgment—in this case, judging the 500 gram object to be heavier. In addition, the evidence shows two other patterns. First, if one stimulus is held constant while the other is changed so that the difference between them becomes greater, the chance of making a mistake reduces: for example, if the heavier weight were increased to 550 grams, the frequency of mistaken judgments would be considerably lower. Second, if the difference between the two stimuli is then held constant at 50 grams but their magnitudes are both increased, the chance of making a mistake increases: that is, if we add 1.5 kilograms to both weights so that they become 2 kilograms and 2.05 kilograms respectively, we should expect many more errors than in the case where 500 grams is being compared with 550 grams.

In other words, although one object *actually* weighs more than the other, the imperfections of human perception/judgment mean that on some occasions, so long as the difference is not too great, the lighter object may be *perceived/judged* as being at least as heavy as the heavier object.

This can be modeled as follows. Denote the actual weights of the two objects by W_1 and W_2 . The individual lifts the first object and mentally registers its weight as $W_1 + \varepsilon_1$, with ε_1 being a random variable representing the degree to which the individual’s perception of weight is labile. Then the second object is lifted and its weight is mentally registered as

⁸Fechner built on the work of E. H. Weber, and some of the relevant literature refers to the “Weber–Fechner law” as an attempt to formalize the relationship between objective and subjective measures.

$W_2 + \varepsilon_2$. The judgment about which object is heavier is then made on the basis of whether $W_1 + \varepsilon_1$ is greater than, equal to, or less than $W_2 + \varepsilon_2$.

As long as ε_1 and ε_2 are independent random errors, the probability of identifying the first object as the heavier can be expressed as $\text{pr}[(W_1 - W_2 + \varepsilon) > 0]$, where ε is a symmetrically distributed random variable with a mean of zero. One interpretation of this is as follows. The true "core" difference is $W_1 - W_2$, and if human perception were perfect, judgment would always correspond with that true difference. But in a world of imperfect judgment it is as if some additional random amount is either added to, or subtracted from, the true difference. If this additional noise happens to operate in the same direction as the true difference or else only partially offsets it, the truly heavier object is correctly identified; but if the random element happens to be a large enough disturbance in the opposite direction, it outweighs the true difference and results in a mistaken judgment.

If the variance of ε is some increasing function of the magnitude of the smaller stimulus, the two main patterns in the data can be accommodated, that is, if we hold the smaller stimulus constant and increase the larger stimulus, the probability of making a mistake will fall; and if we hold the absolute difference between stimuli constant but increase the magnitude of both by the same amount, the probability of making a mistake will increase.

A model of this kind may seem to many economists the obvious one to apply to the data from individual decision experiments: it fits with the econometric convention of a deterministic core combined with a well-behaved disturbance term; and it also appears to be justified by a substantial body of psychophysical research into judgment. So it is not surprising that this form of error model has been popular as the basis for the statistical/econometric analysis of data from individual decision experiments.

For example, Hey and Orme (1994) used a version of the Fechner model to estimate the parameters of various competing core decision theories and to assess which provided the best fit to their data. For each theory in turn, they supposed an individual's true preferences to be determined according to that theory. On that basis, the *true net advantage* of any option f over any other option g can be expressed as $V(f, g)$. For example, taking standard expected utility (EU) theory as the core, $V(f, g)$ is the difference between the expected utilities of f and g computed according to the individual's true von Neumann-Morgenstern utility function, $u(\cdot)$. Alternative theories generate their own values of $V(f, g)$ according to their particular functional forms.

To this core value, Hey and Orme added a Fechnerian noise term, assuming that the actual choice on any particular occasion is made according to the net advantage of f over g , as *perceived* at the moment when the choice is made and given by $V(f, g) + \varepsilon$. In their version of the model, Hey and Orme supposed that the variance of ε was constant across all (f, g) pairs. On that basis, they used standard econometric techniques to compare the relative performance of a number of different core theories.

However, the assumption of constant variance of ε is only one of a number of ways that a Fechnerian error term might be implemented. An obvious alternative, consonant with the psychophysical evidence, might be to model the variance of ε as an increasing function of the magnitude of the stimuli (in this case, that magnitude being the core EU of each option). And there are more elaborate possibilities: the variance might (also) be a function of the complexity of a prospect, and/or may be distributed asymmetrically around the core EU, and so on. We shall discuss some of these possibilities in subsection 7.2.4, after we have outlined the other main candidate model to have emerged from the literature so far.

7.2.3 The Random Preference Model

Instead of supposing an individual to have a single true preference function to which white noise is added, the random preference (RP) approach (see Becker et al. 1963; Loomes and Sugden 1995) supposes it is as if an individual's preferences consist of a *set* of such functions. The intuition here is that any individual's perceptions and judgments are liable to vary to some extent from one moment to another along some spectrum of "states of mind." If we think of each state of mind as being represented by a slightly different preference function, the probability of any one of these functions being applied to the decision in hand is given by the probability of the individual being in the corresponding state of mind at the time when that decision is made. So to say that a particular individual behaves according to a certain "core" theory is to say (i) that the individual's preferences can be represented by some set of functions, all of which are consistent with that theory; and (ii) that for any particular decision task, the individual acts as if she picks one of those functions at random from the set and applies it to the task in question; then (iii) "puts back" that function into the set before picking again at random when tackling another decision (even if it is the identical task encountered another time).

When applied to EU, the RP model supposes an individual's preferences to be represented by some set of von Neumann–Morgenstern utility functions $u(\cdot)$, any one of which may be chosen at random to be applied to a particular decision task. To illustrate, imagine some set of concave increasing functions all calibrated so that the utility of a zero gain is set at zero. For any three payoffs $x_1 > x_2 > x_3 \geq 0$, each function entails $u(x_1) > u(x_2) > u(x_3) \geq 0$, but different functions are liable to entail different subjective judgments about the extent to which x_1 is better than x_2 , x_2 better than x_3 , and so on.

Consider two binary lotteries, as follows. Lottery **A** offers payoff x_1 with probability p and x_3 with probability $1 - p$, while lottery **B** offers x_2 with probability q (where $q > p$) and x_3 with probability $1 - q$. Under EU, **A** will be preferred to (less preferred than) **B** according to whether $[u(x_1) - u(x_2)]/[u(x_2) - u(x_3)]$ is greater than (less than) $[q - p]/p$. Representing the variability of an EU maximizer's judgment in terms of a set of $u(\cdot)$ functions amounts to allowing for that individual's perceptions of how $[u(x_1) - u(x_2)]$ compares with $[u(x_2) - u(x_3)]$ to vary from one state of mind to another. On those occasions where his state of mind is such that he judges $[u(x_1) - u(x_2)]/[u(x_2) - u(x_3)]$ to be greater than $[q - p]/p$ —that is, where it is as if he has selected at random a function $u(\cdot)$ for which that is true—he chooses **A**; alternatively, on other occasions where it is as if he has picked at random a $u(\cdot)$ for which $[u(x_1) - u(x_2)]/[u(x_2) - u(x_3)] < [q - p]/p$, he chooses **B**.

This way of modeling the stochastic nature of preferences is easily extended to tasks such as judging equivalences. Suppose he is asked to state a sure payoff x_C such that he is indifferent between the certainty of that payoff and playing out lottery **A**. Under the RP version of EU, it is as if he picks some $u(\cdot)$ at random and then identifies the payoff x_C such that $[u(x_1) - u(x_C)]/[u(x_C) - u(x_3)] = [1 - p]/p$. Since the value of x_C that satisfies this equation is liable to vary from one $u(\cdot)$ to another, the RP model entails some distribution of x_C derived from the distribution of $u(\cdot)$ that constitutes the individual's stochastic preferences.

7.2.4 Comparing the Fechner and RP Approaches

Chapter 3 discussed the Duhem–Quine thesis (DQT) in some detail. The DQT asserts that it is impossible to test any particular target hypothesis in isolation from (a possibly large number of) auxiliary assumptions. The corollary of this is that the seeming failure of the target hypothesis might in fact be attributable to a failure of one or more of the auxiliary assumptions. Put another way, the implications or predictions of a par-

ticular theory may to some extent depend upon, and perhaps vary with, the nature of the auxiliary assumptions being made.

In this sense, the nature of the stochastic component in people's individual or interactive decision making might be seen as the subject matter of a set of auxiliary assumptions. Thus, what is being tested by any particular experiment or program of experiments is not just one or more "core" theories, but one or more combinations of a core theory with a stochastic specification; and the difficulty raised by the DQT is the problem of knowing exactly what is being rejected if the results of an experiment contradict some implication of a particular "core-plus-error" combination. Is it the core theory that is wrong? Is it the wrong stochastic specification? Or is it both?

The rest of this subsection takes its cue from those questions and considers examples where combining the same core theory either with some form of Fechner model or else with some version of the RP approach may lead to radically different implications. For each case, we discuss what to infer from the available evidence.

We start with what is generally taken to be the most widely and reliably replicated violation of the Independence axiom of EU theory: namely, the form of the Allais paradox now known as the "common ratio effect" (CRE). This effect was introduced in box 2.6 (p. 73), but it will be helpful here to set its key features out more fully. Consider two pairwise choices of the following form:

Choice 1 $R_1 : (x_1, p; x_3, 1 - p)$ versus $S_1 : (x_2, q; x_3, 1 - q)$;

Choice 2 $R_2 : (x_1, \lambda p; x_3, 1 - \lambda p)$ versus $S_2 : (x_2, \lambda q; x_3, 1 - \lambda q)$;

where $x_1 > x_2 > x_3 \geq 0$, $p < q = 1$, and $0 < \lambda < 1$.

The Independence axiom of deterministic EU theory requires that if an EU maximizer prefers the riskier lottery R_1 in Choice 1, she should also prefer the riskier lottery R_2 in Choice 2; or alternatively, if she prefers the safer lottery S_1 in Choice 1, she should also prefer S_2 in Choice 2. Yet it has very often been found that a substantial proportion of any sample either choose R_1 in Choice 1 and S_2 in Choice 2, or else choose S_1 in Choice 1 and R_2 in Choice 2. Moreover, the combination of S_1 and R_2 is much more frequently observed than the opposite "violation" in the form of choosing both R_1 and S_2 . Any such violations are incompatible with the deterministic form of EU; but what if we consider some stochastic form of EU? Might that be capable of accommodating the existence and asymmetric pattern of violations?

Let us start with an RP version of EU. In this case, the answer is straightforward and requires no particular restrictions on the distribution of the $u(\cdot)$ functions. Setting $u(x_3) = 0$, all that needs to be

said is that whatever proportion of an individual's $u(\cdot)$ functions entail $u(x_1)/u(x_2) > q/p$ gives, for that individual, the probability that she will choose R_1 in Choice 1; and exactly the same proportion of that individual's $u(\cdot)$ functions will entail $u(x_1)/u(x_2) > \lambda q/\lambda p$, for all λ , which means the probability that she chooses R_2 in Choice 2 is exactly the same as the probability of choosing R_1 in Choice 1. Let that probability be denoted by α , which may vary from one individual to another. So for any individual, the probability of choosing R_1 and also R_2 is α^2 , while the probability of choosing S_1 and S_2 is $(1 - \alpha)^2$. Of course, if $0 < \alpha < 1$, there are also probabilities of making choices that violate deterministic EU: the individual will choose R_1 and S_2 with probability $\alpha(1 - \alpha)$ and will choose S_1 and R_2 with the same probability. So even without placing any restrictions on the distributions of individuals' $u(\cdot)$ functions, the RP version of EU produces a clear null hypothesis: namely, that there should be no significant asymmetry in the frequencies with which the combinations R_1 & S_2 and S_1 & R_2 are observed.

However, there are now many experimental datasets that reject these implications with a high degree of confidence: as noted above, the common ratio effect has typically involved frequencies of S_1 & R_2 so much greater than frequencies of R_1 & S_2 that the asymmetry is extremely unlikely to have occurred by chance if both combinations are truly equally probable. Thus, if the RP approach is the appropriate way to model the stochastic nature of preferences, then EU is the wrong core theory, and an alternative core is needed. Alternatively, if EU is to be defended as the core, some other stochastic specification must be invoked. Can some form(s) of the Fechner model accommodate these data?

It turns out that combining EU with one relatively simple form of the Fechner model *can* accommodate these asymmetries—although only up to a point. To see this, consider the Fechner model with the assumption made by Hey and Orme (1994) that ε is symmetrical around zero and has constant variance across all pairwise choice problems.

In most reported cases of CRE, the (great) majority of respondents choose the safer lottery S_1 in Choice 1 (especially when S_1 offers the *certainty* of x_2). Under the Fechner model, this implies that for most individuals, $qu(x_2) - pu(x_1) + \varepsilon > 0$: that is, the typical difference between expected utilities is sufficiently large and positive that it is only overturned by a negative ε in a (small) minority of cases.⁹

⁹Of course, there may also be respondents who are risk seeking to the extent that $qu(x_2) - pu(x_1) < 0$; and so long as this difference is not overturned by a sufficiently large positive ε , they will choose their truly preferred option, R_1 .

But now consider the choice between R_2 and S_2 . Here, the difference between expected utilities is $\lambda[qu(x_2) - pu(x_1)]$, so that with λ often taking values as small as or smaller than 0.25, the “true” difference in expected utilities in Choice 2 is only a small fraction of that in Choice 1. With the variance of ε held constant, this means that the true difference is more likely to be overturned in Choice 2: that is, a bigger proportion of the majority who truly prefer S_1 to R_1 will now choose R_2 rather than S_2 , thereby generating S_1 & R_2 observations. And although there will also be a greater likelihood that someone in the minority group who truly prefer R_2 will actually pick S_2 (thereby generating R_1 & S_2 observations), it is easy to imagine that these will be considerably outnumbered by the S_1 & R_2 observations, producing just the kind of asymmetry that has been found in so many studies.

On this account, then, it might seem that the asymmetry is not a violation of EU: on the contrary, the CRE data might be no more than a manifestation of what we should expect from a specification of EUT that allows for a Fechnerian stochastic element in people’s preferences. Could it be that the data have simply been misinterpreted and that this formulation of the theory should be rehabilitated as the core model of decision making under risk? It turns out that it would be premature to jump to that conclusion, for several reasons.

First, although *much* of the existing CRE evidence might be accommodated by this story, not all of it can be. To see why, consider what happens as λ tends toward zero. Under these circumstances, any true difference between expected utilities would also tend toward zero, so that the actual choice made would depend increasingly on the realization of ε . But since ε is symmetrical around zero, this means that in the limit we should expect choices to split 50:50. What we should *not* expect is that the modal preference should actually cross the 50:50 line. And yet there are at least some cases where we observe a substantial majority preferring the safer lottery in Choice 1 but a substantial majority preferring the riskier option in Choice 2, and this switch of modal preference is *not* explicable in terms of this version of the Fechner model. Another result that certainly cannot be accommodated by that model is the case where the majority choose the riskier lottery even in Choice 1 and an even bigger majority choose the riskier option in Choice 2, in contrast to the implication of the model that the Choice 2 split should be closer to 50:50. Such a result is not often reported—mainly, one supposes, because most experimenters set the parameters so as to induce a majority preference for the safer option in Choice 1—but an example can be found in Bateman et al. (2006).

So even within the realm of CRE data, adding on a Fechner noise term with constant variance does not rescue EUT. In addition, such an addition cannot accommodate the "other" version of the Allais paradox now known as the "common consequence effect" (CCE).

An example of the CCE scenario was given in footnote 13 of chapter 4, but it may help to summarize it in more general terms. Here, the two pairwise choices are as follows:

Choice 1 $R_1 : (x_1, p; x_2, r; x_3, 1 - p - r)$ versus $S_1 : (x_2, 1)$,

Choice 2 $R_2 : (x_1, p; x_3, 1 - p)$ versus $S_2 : (x_2, 1 - r; x_3, r)$,

where $x_1 > x_2 > x_3 \geq 0$.

In this case, each of options R_2 and S_2 is produced by taking, respectively, R_1 and S_1 and replacing the r probability of x_2 with an r probability of x_3 . Under EUT, therefore, the expected utilities of R_2 and S_2 are, respectively, lower than those of R_1 and S_1 by $r[u(x_2) - u(x_3)]$. Thus those EU maximizers who truly prefer R_1 to S_1 also truly prefer R_2 to S_2 ; and vice versa. So once again there are two ways of violating EU: either by choosing R_1 in Choice 1 and S_2 in Choice 2; or else by choosing S_1 in Choice 1 and R_2 in Choice 2. Just as with CRE, what has been observed in experiments is that the numbers of people departing from EUT by choosing S_1 & R_2 are substantially greater than the numbers choosing R_1 & S_2 .

The reason why this pattern cannot be accommodated, even in part, by adding a Fechner term with constant variance is that since the expected utilities of R_2 and S_2 are, respectively, lower than those of R_1 and S_1 by $r[u(x_2) - u(x_3)]$, the true difference between the alternatives is exactly the same for each choice. If ε has the same distribution in both cases, the chance of a true R_1 -preferer actually picking S_1 as a result of noise is exactly the same as the chance of a true R_2 -preferer picking S_2 ; and conversely. Thus under this model the appropriate null hypothesis for the CCE is that both forms of violation are equally likely to occur—a null that would clearly be rejected by the bulk of the experimental evidence.

Of course, the assumption that ε has constant variance is only one possible auxiliary assumption. Might some other specification do better?

As noted at the end of subsection 7.2.2, it might be more in keeping with the psychophysical origins of the Fechner model to suppose that the variance of ε is positively correlated with the magnitude of the stimuli, which in this case might mean that the variance increases as the magnitude of expected utility increases. But such a model does even less well in accommodating the CRE and CCE data. In both cases, it would entail the variance of ε being smaller in Choice 2 than in Choice 1. For the CRE, this

Note: must double check this cross-reference late in the production process, when all box footnotes are in place.

would be liable to ameliorate any tendency to switch as the probabilities of the positive payoffs are scaled down.¹⁰ For the CCE, it would tend to produce the *opposite* patterns to the ones observed: with the variance of ε greater in Choice 1 but the true difference between expected utilities the same for both pairwise choices, the implication is that the split would be driven more by noise—and hence would be closer to 50:50—in Choice 1 than in Choice 2, which is quite contrary to the evidence.

Another possibility mentioned in subsection 7.2.2 is that the variance of ε is some function of complexity: options involving several payoffs may make heavier cognitive demands and be more prone to noise than those involving just two payoffs; and/or options involving one positive and one negative payoff, with respective probabilities related to a roulette wheel and therefore expressed in multiples of $1/36$, may be more taxing to evaluate than ones that involve just a positive and a zero payoff, with probabilities expressed as multiples of 0.05 . However, as yet there does not seem to be any well-developed theory of complexity that enables us to make the variance of ε a function of some set of measurable independent variables.¹¹

A somewhat different way of trying to accommodate CRE and CCE patterns within a stochastic form of EU has been proposed by Blavatskyy (2007). The key idea here is to take an initially symmetric distribution of ε around the true EU of a lottery but then truncate and redistribute it so that all realizations of $EU + \varepsilon$ lie within bounds set by the highest and lowest payoffs. In conjunction with some rather particular assumptions—that the true EU of R_1 is a little higher than the true EU of S_1 and that the true EU of R_1 is sufficiently close to $u(x_1)$ —Blavatskyy shows how both the CRE and the CCE might be compatible with this model. However, as we shall see below, this model is unable to account for other well-attested phenomena; and since it fails in ways that could be regarded as direct tests of its basic premises, its potential as a general model of stochastic EU would appear to be limited.

¹⁰Much would depend on the particular relationship between the magnitude of the stimuli and the variance of ε , but if, for example, the reduction in the variance were such that the true expected utility difference accounted for a constant proportion of the distribution of ε , there would be no systematic tendency whatsoever to switch from safer to riskier, or vice versa.

¹¹Even in the absence of such a theory, there have been some attempts to explore other factors that might be influential. Buschena and Zilberman (2000), for example, considered three different variables that might affect the variance of ε . One specification made the variance a function of the “true” difference between the values of the alternatives (as estimated according to whichever core theory was being examined); the second specification allowed the variance to increase with the average number of outcomes for which the two lotteries had positive probabilities; the third allowed the variance to be affected by the area between the lotteries’ cumulative distribution functions over the range of outcomes.

Of course, there are many other variants of a Fechnerian error model that we might consider in conjunction with different implications of an EU core—or indeed, different implications of any one of the many non-EU alternative core theories that have been proposed during the past three decades. But at every turn we are liable to run up against the DQT difficulty of trying to decide how far any divergence between model and data might be attributable to one assumption or another. So if our primary concern is to distinguish between one stochastic specification and another, a more efficient approach might be to take some principle or axiom that is common to many core theories and that commands very wide, if not universal, acceptance and then use data about that principle to try to discriminate between different stochastic specifications.

One such principle—indeed, possibly the only such principle—is respect for First-Order Stochastic Dominance (FOSD), as described earlier in box 2.5 (p. 71). What has become apparent in the course of hundreds of experiments examining individual decision making under risk is that although almost every *other* axiom or basic postulate about rational choice is liable to be violated in seemingly systematic ways, *transparent* FOSD appears to be the exception: so long as the dominance relation is transparent, it is respected by the overwhelming majority of participants in experiments.¹²

However, such respect for all forms of transparent FOSD is at odds with the Fechner model. To see this, recall the evidence from Loomes and Sugden (1998), cited earlier, where forty choices *not* involving dominance were each presented to respondents on two occasions within the same experimental session, and where the choice on the second occasion was different from that on the first in 18.4% of cases. Scattered among those forty choices were another five pairs where one lottery dominated the other. In each of these pairs, the lotteries were really quite similar to each other, mostly with one offering a 0.05 higher chance of £20 or £30 than the other and a corresponding 0.05 lower chance of 0. Thus each of these five pairs involved differences between expected values in the region of £1 or £1.50: that is, considerably smaller differences than in most of the other forty choices. So although differences in expected values may only be a rough proxy for $V(\mathbf{f}, \mathbf{g})$, any variant of the Fechner model discussed above can reasonably be expected to entail error rates at least as high as observed in the majority of the other forty choices.¹³

¹² However, when dominance is disguised, it is possible to induce substantial rates of violation (see, for example, Tversky and Kahneman 1986 and Charness et al. 2007).

¹³ This is clearly true for the Hey and Orme (1994) constant-variance version; and a model that allows the variance of ε to vary with the magnitudes of \mathbf{f} and \mathbf{g} would give much the same prediction, since the difference in their magnitudes is small. By the stan-

But this was not the case at all: in fact, out of a total of 920 observations (ninety-two respondents each making five choices on two occasions), FOSD was violated in just thirteen cases—a rate of less than 1.5%.

This suggests that the Fechner model, which works well in the context of simple physical stimuli, is liable to greatly overpredict violations of FOSD if transplanted into the context of decision making under risk, provided that dominance is relatively easy to detect. Arguably, this is because the two contexts are crucially different, in the following way.

When a respondent is judging the relative heaviness of objects, what matters is the “resultant weight” of each object: that is, the overall force exerted by gravity on each of them. The variability in judgment comes solely from the respondent’s perception of the muscular force he has to exert to counteract gravity. For this model to carry through to choices between lotteries, it would have to be the case that the respondent processes the task by evaluating each lottery as if in isolation, so that only the “resultant weight” of each lottery, taken individually, matters.

This *might* be the case when a certainty equivalent is elicited for each lottery separately: under those conditions, it would not be too surprising to find some of the certainty equivalents elicited for the dominated lottery on some occasions being higher than some of the certainty equivalents elicited for the dominating lottery on other occasions. For example, Cubitt et al. (2004) reported an experiment where respondents were asked to value a number of lotteries, one of which offered a 0.36 chance of £7 and a 0.64 chance of zero, while another offered a 0.41 chance of £7 and a 0.59 chance of zero. When these were valued separately, 36 out of 230 respondents (15.7%) gave a higher value to the dominated alternative. But when those same respondents were asked to make a straight choice between those two lotteries, only 7 (3%) of them chose the dominated alternative. That is, when the two lotteries were compared *directly*, so that the superiority of one over the other was quite transparent, the error rate was greatly reduced. This is contrary to the idea that it is only the “resultant weight” of each lottery that matters, since those resultant weights were the same for both valuation and choice. That in turn casts doubt upon any error story that supposes that a choice between two alternatives can be modeled as if each is evaluated separately and subject to independent error prior to a decision being made.¹⁴

dards of the Buschena and Zilberman (2000) model, f and g were roughly equally complex, and no less complex than other lotteries in the experiment. And since f and g had the same upper and lower bounds and their EUs were close to each other but often not close to either bound, Blavatsky’s (2007) model would also predict violations of FOSD by a substantial minority of respondents.

¹⁴In fact, this is precisely the supposition that Blavatsky (2007) makes. His model amounts to saying that for any nondegenerate monetary lottery, the distribution of $EU + \varepsilon$

But is RP a better alternative? Consider first how RP fares in relation to the kind of behavior observed by Cubitt et al. (2004) and cited above. As seen in subsection 7.2.3, RP entails that an individual has a distribution of certainty equivalents for a lottery offering a 0.36 chance of £7 and also has a distribution of certainty equivalents for a lottery offering a 0.41 chance of £7. For any set of $u(\cdot)$, the distribution of certainty equivalents will be shifted to the right as the probability of the positive payoff increases, but there is liable to be some overlap of the two distributions—at least so long as the probabilities are not *too* different. So if we model the certainty equivalent stated on one particular occasion for (£7, 0.36) as a value drawn at random from one distribution, and the certainty equivalent elicited on another occasion for (£7, 0.41) as a value drawn at random from another distribution somewhat to the right but overlapping with the first, we can accommodate cases where a nontrivial minority (in the case in question, 15.7%) give a higher value to the dominated lottery.

By contrast, RP does not allow *any* violation of FOSD in a direct choice between the two lotteries. Let the dominant lottery be labeled D and call the dominated lottery E . For E to be chosen over D would require $[u(7) - u(7)]/[u(7) - u(0)]$ to be greater than $[q - p]/p$; but for every $u(\cdot)$, $[u(7) - u(7)]/[u(7) - u(0)]$ must be zero while $[q - p]/p$ is strictly positive, so that there is a zero probability of picking the dominated lottery in a straight choice between the two,¹⁵ except as a result of some “tremble” due to occasional lapse of attention, misreading, etc. However, as noted earlier, the RP model does not exclude the possibility of some additional source of noise/error of the kind captured by trembles, so RP plus a small tremble term may be a plausible way of accounting for the low but positive rate of violations of transparent FOSD actually

will map directly to a distribution over the sure sums of money between the highest and lowest payoffs offered by the lottery. Thus, making a choice between two nondegenerate lotteries is supposed by this model to be exactly the same as drawing a sure value independently from each lottery’s distribution, then comparing the two and choosing whichever lottery happens on that occasion to be associated with the higher value. It is therefore a fundamental implication of the Blavatsky model that the pattern of preference inferred from comparing the two valuations should be no different from the pattern observed via direct choice. However, this implication is clearly rejected by the case reported by Cubitt et al. (2004). More generally, it is rejected by the very large number of “preference reversal” datasets (about which more later) showing pronounced and systematic differences between preferences inferred from valuations and those revealed by straight choices.

¹⁵ Although the notation here is most readily associated with EU, the argument carries through much more generally. Instead of a von Neumann-Morgenstern $u(\cdot)$ function, we might consider any subjective value function $v(\cdot)$, requiring only that it is a non-decreasing function of wealth; and instead of using the “objective” probabilities p and q as weights, we might allow any transformation so long as it does not entail assigning negative weights to positive probability differences.

observed in many experiments—including the one conducted by Cubitt et al. (2004). But according to RP, any such rate can be expected to be considerably lower than the propensity to give a higher certainty equivalent to E than to D when those certainty equivalents are elicited on different occasions; and this, too, accords with the evidence.

Thus—at least when considered in the context of one particular principle that is common to many core theories—it would appear that the RP framework is superior to the Fechner approach for modeling noise/error in decisions involving risky prospects. Of course, we should be wary of jumping too confidently to the conclusion that RP is *the* right model on the strength of the current rather limited evidence base. But *if* RP is a more appropriate way of allowing for the stochastic component in people's decision making, the implication is clear, since applying RP to EU generates null hypotheses that are strongly rejected in large numbers of experiments. On *this* reading, then, we would have to conclude that it is the EU core rather than the auxiliary assumptions that fails.

This raises a number of questions for future research, including:

- (1) *If* RP is the appropriate model and EU is the wrong core, how should RP be applied to alternative core theories in order to test them and to discriminate between their respective claims? In particular, different restrictions on the distributions of parameters central to the various core theories may well have different implications for which patterns of behavior are consistent with the core theory, so which restrictions should be adopted? In the discussion above, about how RP might be applied to EU, it was supposed that the variability of judgment from one decision to another is exclusively about the relative subjective values of the payoffs, that is, about how $[u(x_1) - u(x_2)]/[u(x_2) - u(x_3)]$ varies from one judgment to another; and it was supposed that choice involved comparing the realization of $[u(x_1) - u(x_2)]/[u(x_2) - u(x_3)]$ with the ratio $[q - p]/p$, which was taken as given and as always perceived as taking its “objective” value. However, a number of alternative core theories involve some transformation of “objective” probabilities into subjective decision weights. Should an RP version of such theories require not only that $[u(x_1) - u(x_2)]/[u(x_2) - u(x_3)]$ —or its counterpart in the alternative theory—is determined on any particular occasion as if on the basis of a random draw from some set of $u(\cdot)$, but also that $[q - p]/p$ is determined as if on the basis of some independent random draw from some set of probability transformation functions?

- (2) If both the EU core and the essence of the Fechner approach are inconsistent with the data from individual decision experiments, can they provide a credible basis for explaining other experimental data—for example, those generated in experimental games and markets? And if not, could an RP specification of some alternative core theory be used to analyze and organize those data?
- (3) Or is there some other way of modeling the nature of people's preferences and beliefs that is different—and perhaps radically so—from any of those discussed so far?

To have some chance of answering the first of these questions, experimenters would need to undertake a substantial program of research. However, before discussing what such a program of research might entail, we turn to the second question and, in the next section, consider the additional issues raised by data from experimental games.

7.3 "Noise" in Experimental Games

As discussed in chapter 3, in standard presentations of (complete information) game theory, players are assumed to know their own and others' payoffs contingent upon all possible combinations of their own and others' strategies. Moreover, it is conventionally assumed that they know all those payoffs in (von Neumann–Morgenstern) utility terms. The usual assumption is that any player is only interested in others' utilities to the extent that they inform him about the probabilities of the strategies that the other player(s) might select—information that he can then take into account when formulating his own strategy.

When games are implemented experimentally, payoffs are typically expressed in the form of sums of money. This may be easier for participants to understand and it provides straightforward incentives, but in the light of the discussion above, it also means that participants may not know the precise utility of each payoff to each potential recipient, which is what standard theory assumes. Indeed, the lesson from individual decision experiments is that most individuals do not know their *own* utilities with precision, in which case it seems plausible to suppose that they have even less precise ideas about how payoffs map to utilities for other players.

Consider a normal form presentation of a 2×2 game where the row player (Row) has to decide between U(p) and D(own) and the column player (Col) chooses between L(eft) and R(ight), with all payoffs in

Table 7.1. A normal form representation.

	Left	Right
Up	4, w	0, x
Down	0, y	1, z

Table 7.2. A mixed strategy game.

	Left	Right
Up	4, 2	0, 3
Down	0, 4	1, 1

some money currency. In the example shown in table 7.1, Row's payoffs are given as specific numbers, while Col's are, for the moment, left unspecified as w , x , y , and z .

If payoffs were utilities and both players knew each other to be EU maximizers, Row would still need to figure out the probability that Col will play L or R. In some games, that might be supposed to be quite easy. For example, if $w > x$ and $y > z$, Col would be expected to see that L dominates R,¹⁶ and she would therefore be expected to play L for sure; and realizing this, Row would play U. Similarly, if $w < x$ and $y < z$, Col would identify R as her dominant strategy and play it for sure; and knowing that, Row would choose D.

But there are many patterns of w , x , y , and z that do not lead to such straightforward conclusions, even if payoffs are utilities and players are both EU maximizers. For example, consider the case shown in table 7.2 where neither L nor R dominates the other.

There is no pure strategy Nash equilibrium in this game: if Row were to play U, Col's best response would be R, to which Row's best response would be D, to which Col's best response would be L, to which Row's best response would be U, and so on. The unique Nash equilibrium is in mixed strategies, with each player playing their pure strategies with probabilities that make the other player indifferent, so that in equilibrium each player is willing to randomize in whatever way satisfies the "mutual indifference" requirement. While not all game theorists would agree that such Nash equilibria are appropriate as predictions of play,

¹⁶Notice that we are here using the game-theoretic notion of dominance. To avoid possible confusion and to distinguish this usage from First-Order Stochastic Dominance in cases where probabilities are known, the latter will be referred to in terms of *stochastic dominance* and/or FOSD.

they are often taken to constitute the maintained hypothesis in experimental tests of standard game theory, when there is no Nash equilibrium in pure strategies.

In the example shown in table 7.2, Col's Nash equilibrium mixed strategy is to play L with 0.2 probability and R with 0.8 probability so that Row gets the same expected utility of 0.8 from both U and D. Row therefore does not mind which mix he adopts and is willing to play U with probability 0.75 and D with probability 0.25, which makes Col indifferent between L and R (both yielding an EU of 2.5) so that she remains content with her own mixed strategy.

However, when such games are being implemented experimentally, noise may arise from several sources.

First, as noted above, in the great majority of experimental games, payoffs are sums of money. Such cases are liable to entail the kind of noise discussed in section 7.2: namely, the noisy translation of money payoffs into utilities. Even if Row were told the probabilities with which Col would play L and R—let us call them p and $1 - p$ respectively—and if Row were a stochastic EU maximizer, he would have to judge whether $p[u(4) - u(0)]$ was greater or less than $(1 - p)[u(1) - u(0)]$, and for some (range of) $p < 0.5$, his stochastic preferences might mean that there were some moments when he would perceive U to have a higher expected utility than D and other moments when he would perceive D to be the better option. Moreover, the proportions of U choices and D choices would be liable to vary with p : when p is close to 0.5, U would very likely be judged better than D, but progressively lower values of p diminish the perceived value of U and raise the perceived value of D so that U becomes less and less likely to be judged preferable to D.

Much the same is also true for Col's decisions, of course: even if she knew the chances of Row playing U and D—call these q and $1 - q$ respectively—there are liable to be values of q such that Col sometimes perceives $qu(2) + (1 - q)u(4)$ to be greater than $qu(3) + (1 - q)u(1)$, and sometimes makes the opposite judgment. And the proportion of times L is judged better than R could be expected to fall as q rises.

So just allowing for the kind of noise suggested by individual decision experiments has the effect of greatly complicating experimental game behavior. Even under the assumption that players have EU preferences, the probability that each player will play a particular strategy itself has a stochastic element. In the next two subsections we shall consider two possible ways of incorporating stochastic judgment into the modeling of strategic behavior.

Table 7.3. Asymmetric Matching Pennies payoffs.

	Left	Right
Up	4, 0	0, 1
Down	0, 1	1, 0

7.3.1 Quantal Response Equilibrium

A technically neat way of modeling this seemingly untidy problem is McKelvey and Palfrey's (1995, 1998) notion of a *quantal response equilibrium* (QRE). To illustrate how the QRE idea works, and examine its strengths and possible limitations, consider a case where the money values of w , x , y , and z in table 7.1 are set so as to produce an "asymmetric matching pennies" game, as in table 7.3.

Again, there is no Nash equilibrium in pure strategies in this game. Rather, the conventional approach is to suppose that players' utilities are proportional to their money payoffs and to identify the mixed strategy equilibrium. It entails Row playing U and D with equal probability while Col plays L and R with probabilities 0.2 and 0.8 respectively.¹⁷ If a sample of participants behave as if they are implementing such mixed strategies, we should expect to observe those strategies being played with the corresponding frequencies.

However, there is evidence from experiments that under such circumstances, Row players choose U significantly more than 50% of the time (see Ochs 1995; Goeree et al. 2003). QRE attempts to provide an account for such data by adding a Fechnerian stochastic element to players' behavior.

Consider Row's decision. If Row's behavior has a stochastic component of the Fechnerian kind, the probability of choosing U from the pair (U, D) is given by $\text{pr}[V(U, D) + \varepsilon > 0]$. Thus Row's probability of choosing U or D depends on an interaction between her belief about Col's strategy and the distribution of ε .

A parallel story applies to Col. If she believes that there is a better than 0.5 chance that Row will play U, the standard deterministic model entails her playing R with probability 1. But if the probability of playing

¹⁷The assumption that utilities are strictly proportional to money payoffs could be relaxed by assuming instead that players operate according to some nonlinear utility function. If we set $u(0) = 0$ and if players are risk averse, so that $u(4) < 0.25u(1)$, Col would need to play L with probability greater than 0.2 in order to make Row indifferent between U and D; however, since Col's payoffs are 0 or 1 whichever strategy she plays, Row would still need to play both U and D with equal probability in order to make Col indifferent between L and R.

R is given by $\text{pr}[V(R,L) + \varepsilon > 0]$, then there is some chance that she could play L.

On this basis, an equilibrium occurs when players' beliefs and behavior are mutually compatible, so that there exists some pair of probabilities (p^*, q^*) such that Row's choice satisfies $\text{pr}(U) = q^*$ when Row believes that $\text{pr}(L) = p^*$, and Col's choice satisfies $\text{pr}(L) = p^*$ when Col believes that Row's $\text{pr}(U) = q^*$.

Logit equilibrium is a particular case of QRE where the probabilities of choosing the different pure strategies are proportional to exponential functions of their expected utilities. Goeree et al. (2003) have analyzed a number of datasets on the basis of assuming that the probability of playing strategy i when there are two strategies i and j , denoted by P_{ij} , can be expressed as

$$P_{ij} = \exp(EU_i / \mu) / [\exp(EU_i / \mu) + \exp(EU_j / \mu)],$$

where μ may be thought of as an "error variance" parameter, with higher values of μ giving greater weight to the stochastic component at the expense of the true EU difference.¹⁸ Even under the restrictive assumption of risk neutrality, it appears that the logit equilibrium approach can accommodate many experimental datasets much better than deterministic Nash equilibrium.¹⁹

However, although it achieves a better fit to the data from a number of experiments, there is still a question about whether this is a plausible equilibrium concept. The main difficulty, as with the Fechner model when it is applied to individual decisions, is that, given the equilibrium probabilities, it appears to entail high rates of violation of FOSD.

For example, consider the game in table 7.3 with $\mu = 1$ and risk-neutral players. Here, the logit equilibrium is where $q^* \approx 0.722$ and $p^* \approx 0.391$. But is this a credible equilibrium? For this equilibrium to hold, it requires that if faced with the choice between choosing L and getting a 0.278 chance of 1 or choosing R and getting a 0.722 chance of 1, Col will choose L on nearly 40% of occasions. Yet when seen in this way, L is clearly stochastically dominated by R, and as seen earlier, when the positive payoff is the same and when one lottery offers a better chance of that payoff than another, the stochastically dominated alternative is rarely chosen, even when the probability difference is *much* smaller (0.05) than in the present example.

¹⁸In the limit, as μ goes to infinity, the choice of pure strategy becomes entirely random, with each one likely to be chosen with equal probability.

¹⁹With the additional degree of freedom provided by allowing for subjects' risk aversion, even better fits of even more datasets can be achieved (see Goeree et al. 2003).

How is it, then, that QRE can entail the stochastically dominated strategy being chosen nearly 40% of the time? The answer is that the QRE approach, like the Fechner model in individual decision making, assumes that each strategy is evaluated separately, and that each evaluation involves an independent draw of ε from the noise distribution. Setting $u(x) = x$, the QRE approach supposes that L is evaluated in utility terms as $0.278 + \varepsilon_L$, where ε_L denotes a random draw from the distribution of ε when L is being evaluated; meanwhile, R is evaluated separately as $0.722 + \varepsilon_R$. When $\mu = 1$, $\varepsilon_L - \varepsilon_R > 0.722 - 0.278$ on nearly 40% of the occasions when both evaluations are undertaken independently, so that on these occasions $0.278 + \varepsilon_L > 0.722 + \varepsilon_R$, and thus L is chosen even though when the two are considered together L is clearly stochastically dominated by R, given the equilibrium probabilities of Row's play.

In short, QRE achieves a better fit than Nash equilibrium to some of the data from experimental games, but it does so by making the same contestable assumption that the Fechner model makes when applied to individual decisions under uncertainty: namely, that players behave as if they evaluate the EU of each strategy separately with an independent ε , and are thereby liable to choose stochastically dominated strategies to an extent that seems implausible.²⁰ How different would things be if we used RP rather than the Fechnerian model?

7.3.2 Random Preference in Experimental Games

Consider first the game shown in table 7.2, starting with the view from Row's perspective. If the probability of Col playing L is p , then Row is facing the choice between $U = pu(4) + (1-p)u(0)$ and $D = pu(0) + (1-p)u(1)$. The probability of Row choosing U depends on the probability that she picks a $u(\cdot)$ from her notional set of functions such that $[u(4) - u(0)]/[u(1) - u(0)] > (1-p)/p$.

Of course, by contrast with individual risky choice problems, Row is not told what value p will take: that is determined by Col. However, for each possible value that p might take there will be some corresponding probability q that Row will choose U. For all increasing functions $u(\cdot)$, it

²⁰ Another possible concern relates to the evidence that the value of μ varies a good deal from one type of game to another and, even within the same game type, from one particular experiment to another, even after controlling for scale effects. When commenting on this in relation to different "matching pennies" datasets, Goeree et al. (2003) conjecture that other things that could account for such differences might include factors such as different subject pools and procedures. However, it is not obvious why different subject pools should exhibit different values of μ while their estimated risk aversion parameters are very similar. On the other hand, it might be argued that (scale effects aside) variations in μ across different game types/procedures may to some extent be explicable in terms of some measure(s) of the difficulty or complexity of the different games.

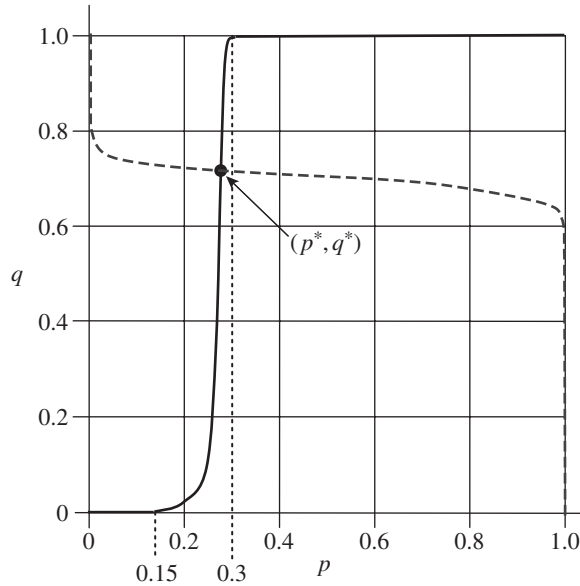


Figure 7.1. An RP equilibrium analysis.

must be the case that $[u(4) - u(0)]/[u(1) - u(0)] > 1$, so we can be sure that U will be chosen for all $p \geq 0.5$; but for at least some $0.5 > p > 0$, we may expect q to fall as p falls.

To illustrate, consider a Row player whose most concave $u(\cdot)$ is such that $[u(4) - u(0)] = 2\frac{1}{3}x[u(1) - u(0)]$, while at the other end of the set his most convex $u(\cdot)$ entails $[u(4) - u(0)] = 5\frac{2}{3}x[u(1) - u(0)]$. For any $p > 0.3$, even his most concave $u(\cdot)$ gives a strict preference for U, so $q = 1$ for all $p > 0.3$. However, once p falls below the 0.3 threshold (where Row is indifferent between U and D when his most concave $u(\cdot)$ is applied), there will be more and more $u(\cdot)$ that entail choosing D, so that q falls. However, once p drops below 0.15, Row will choose D even when his most convex $u(\cdot)$ is applied—which means that for all $p < 0.15$, $q = 0$. Thus for any set of $u(\cdot)$, there will exist some function showing q contingent on p . In figure 7.1, a function of this kind is shown by the solid line.

Correspondingly, from Col's perspective the probability of choosing L represents the proportion of Col's $u(\cdot)$ functions such that $[u(4) - u(1)]/[u(3) - u(2)] > q/(1 - q)$. Suppose Col's most concave $u(\cdot)$ entails $[u(4) - u(1)] = 1\frac{1}{2}x[u(3) - u(2)]$, while her most convex $u(\cdot)$ function gives $[u(4) - u(1)] = 4x[u(3) - u(2)]$. It follows from this that for all $q < 0.6$, L will certainly be chosen (i.e., $p = 1$). As q rises above 0.6, there will be more and more $u(\cdot)$ for which $[u(4) - u(1)]/[u(3) - u(2)]$ falls

short of $q/(1 - q)$, so that R is chosen rather than L and p falls; and once q exceeds 0.8, L will never be chosen, so that $p = 0$ over this range. This relationship between p and q is represented by the dashed line in figure 7.1.

On the assumption that both functions are continuous and weakly monotonic, there will be a unique point at which the two functions will intersect: this point is the pair of probabilities (p^*, q^*) such that Row's $\text{pr}(U) = q^*$ when Row believes Col's $\text{pr}(L) = p^*$, and Col's $\text{pr}(L) = p^*$ when Col believes Row's $\text{pr}(U) = q^*$. This is the RP analogue to QRE for this game.

What makes the RP analysis different from the Fechnerian QRE, however, is that RP builds in respect for stochastic dominance. In the example above, RP entails that Col will *never* choose R when $q \leq 0.5$, whereas QRE permits the possibility that the perceived EU of R (i.e., $qu(3) + (1 - q)u(1) + \varepsilon_R$) may be greater than the perceived EU of L (i.e., $(1 - q)u(4) + qu(2) + \varepsilon_L$) even when $q \leq 0.5$, since there is some probability that $\varepsilon_R - \varepsilon_L$ may be positive enough to outweigh the "true EU" difference between L and R.

This difference becomes especially important in cases such as the matching pennies game in table 7.3. Here, RP entails that Col will choose R or L according to whether $q[u(1) - u(0)]$ is greater than or less than $(1 - q)[u(1) - u(0)]$. In other words, under RP the relationship showing p as a function of q is that $p = 0$ for all $q > 0.5$ and $p = 1$ for all $q < 0.5$, while p can take any value (Col is indifferent between all probability mixes of L and R) when $q = 0.5$. This rules out an equilibrium of the kind permitted by QRE where q^* deviates from 0.5. Thus, RP does not allow any violation of stochastic dominance to arise from the equilibrium concept. So *if* RP is the appropriate way of modeling noise in decision making, the data from matching pennies experiments would reject conventional core theory, represented by the requirement that beliefs and behavior conform probabilistically. As with individual choice, the conclusions one should draw from the data about core theory seem to depend on assumptions about stochastic specification.

In the case of games, conventional core theory may be "saved" by QRE—but only if we are prepared to accept an equilibrium concept that allows substantial violations of stochastic dominance, entailing a model of individual decision making that is not well supported by individual choice data. It remains to be established whether more direct tests of QRE will identify other grounds for concern—for example, values of μ varying inexplicably or implausibly under controlled conditions. As in the case of individual decision experiments, there is still much work to be done to explore the possible ways in which noise/error/imprecision

might be modeled and to examine how far data that appear to violate core principles may or may not be accommodated by particular “auxiliary” assumptions—if that is what stochastic models are.²¹

Moreover, if it appears that no plausible stochastic specification of conventional core theory can be reconciled with the data, we shall have to turn our attention to ways in which alternative core theories might be specified stochastically and investigated experimentally. The next subsection indicates some of the issues arising from certain recent developments in modeling behavior in experimental games that take a rather different approach.

7.3.3 Alternative Models of Strategic Behavior

In the previous two subsections, QRE and RP were discussed in conjunction with “standard” core assumptions: in particular, that players are EU maximizers (albeit noisy ones) who believe all other players to be essentially the same as they are. Different players may have differing attitudes to risk and perhaps different degrees of noisiness, but in key respects they are assumed to conform with the standard view of rational players: that is, their judged utility comes only from their own payoffs, and they are interested in others’ payoffs only to the extent that this allows them to form correct beliefs about the probabilities with which the other player(s) will choose their strategies.

However, these core assumptions might be modified in various, more or less radical, ways, and such modifications may be offered as explanations for many of the seemingly robust and systematic deviations from standard predictions. For example, it may be that when respondents are presented with information not only about their own monetary payoffs but also about the other players’ monetary payoffs, they reinterpret their true payoffs to include an additional element due to “social” or “interpersonal” considerations—some attitude to fairness, perhaps, or a degree of envy. Thus, a payoff structure that appears to entail a dominant strategy when analyzed from the perspective of unmodified self-interest may not be seen as such by some players inclined to bring interpersonal considerations to bear (Rabin 1993).

The issue of how such interpersonal considerations might complicate the implementation of experimental games was discussed in chapter 3. Such considerations may make it much harder for each player to judge

²¹If some degree of variability/imprecision/noise is *intrinsic* to human information processing, judgment, and decision processes, it is not obvious that we can so easily distinguish between what are “core” and what are “auxiliary” assumptions; so such a distinction is made subject to that caveat.

Table 7.4. A dominance-solvable game.

	Left	Right
Up	3, 10	2, 2
Down	2, 10	0, 0

the utility value placed by other players on the money payoffs and therefore make it difficult to judge the chances of other players selecting particular strategies—or, indeed, to know exactly what game is “really” being played.

To illustrate, consider the game shown in table 7.4, where the figure gives monetary payoffs from the different strategy combinations.

Viewed conventionally, this game is straightforward: for Row, U dominates D; for Col, L dominates R; and so the equilibrium involves Row playing U and Col playing L and the two players receiving payoffs of 3 and 10 respectively.

But suppose Row is averse to the payoff inequality this outcome entails: so much so, in fact, that the subjective value of a nominal payoff of 3 under these circumstances is perceived by him to be lower than receiving a payoff of 2 when Col gets a payoff of 1. That is to say, if Row anticipates that Col will play her conventionally dominant strategy L, Row would prefer the joint outcome (2, 1) to the joint outcome (3, 10) and will therefore play D. Such a result, were it to account for $x\%$ of all observations, might therefore not be due to Fechnerian noise large enough to overwhelm the true value difference between U and D on $x\%$ of occasions but could reflect $x\%$ of the sample population having more than the relevant threshold level of aversion to being on the wrong end of an unequal outcome.

And then what if column players were alert to the existence and prevalence of such people? If Col were only interested in her own payoffs, she would continue to play her conventionally dominant strategy L, since even just 1 is better than 0. But if she were to feel resentful of Row deliberately trying to deprive her of the 10 she might have had, she might feel a desire to retaliate: rather than allow Row to reduce her payoff to 1 while still getting 2 himself, she might prefer that they both end up with 0; in which case, her best response to D (when she perceives D to have been deliberately chosen for the motives described) would be R. On this basis, L would not be regarded by Col as dominating R.

What this example illustrates is that if players' values are not simply functions of their own nominal payoffs but also take account of *relative* payoffs and/or the interpersonal motivations of other players, each

player's computation of the expected utility of each strategy may involve (a) modifications of their own payoffs and (b) judgments about the probabilities attached to the other player's strategies that depend in part on some estimate of the likelihood that the other player has motivations beyond maximizing the expected value of nominal payoffs and on some assessment of what those other motivations might be.

Under these conditions, even if there is no "noise" of the Fechnerian or RP kind, there would still be patterns of play that would depart from conventional predictions and that *might appear* to be due to noise of some sort. Moreover, there may now be additional sources of "genuine" (Fechnerian or RP) noise. First, not only may an individual's judgment about the relative utility differences between payoffs vary from one occasion to another, but also her judgment about the utility of receiving x_i when the other player is receiving x_j may vary from one occasion to another. So too might her judgment about the proportion of other players who are purely interested in their own payoffs, as opposed to those who take account of relativities and other interpersonal motivations—a source of variability somewhat analogous to the possibility that in "games against nature," objective probabilities may be transformed differently into subjective decision weights on different occasions.

Allowing the possibility that different players may have different propensities to modify nominal payoffs to take account of comparisons with opponents' payoffs and motivations may seem to be complicated enough. But thus far we have continued to suppose that all players, whatever their personal and interpersonal attitudes might be, are capable of the sort of sophisticated reasoning that is characteristic of conventional game theory.

However, when trying to interpret data from experimental games, a further complicating factor is the possibility that many participants in such experiments either themselves have limited capacities for reasoning according to standard game-theoretic logic and/or may believe that other players have such limitations. For example, although participants in individual decision experiments rarely fail to respect transparent FOSD, solutions involving some element of successive elimination of dominated strategies may require several steps of reasoning combined with a belief that the other player can be relied upon to reason similarly and act accordingly. But, in practice, players may not (all) reason in that way and/or may not be totally confident that others do.

Thus, instead of assuming that all players have equal powers of reasoning about each others' strategic choices, an alternative strategy for trying to accommodate the experimental data involves allowing for different "types" of players, characterized by different degrees of sophisti-

cation. The last decade has seen considerable interest in modeling this heterogeneity and fitting such models to the data (see, for example, Stahl and Wilson 1995; Costa-Gomez et al. 2001; Camerer et al. 2004a). Essentially, the modeling strategy adopted by these and other authors involves positing different “levels” or “steps” of sophistication, with step 0 being the most basic and least “reasoned,” with players on higher steps able to take account of the existence and behavior of others on lower levels.

For example, Camerer et al. (2004a, p. 863) propose the idea of a *cognitive hierarchy* (CH) in which “step k ” players

assume that their opponents are distributed, according to a normalized Poisson distribution, from step 0 to step $k - 1$: that is, they accurately predict the relative frequencies of players doing fewer steps of thinking, but ignore the possibility that some players may be doing as much or more.

Camerer et al. (2004a)

Essentially, step 0 (hereafter, S0) players are modeled as behaving at random and thus are assumed to play each available strategy with equal probability; step 1 (S1) players are assumed to act so as to maximize their expected payoff on the supposition that all other players are S0; step 2 (S2) players maximize their expected payoffs on the assumption that the rest of the population is some mix of S0 and S1; and so on. Camerer et al. find such a model to be a parsimonious way of organizing a large amount of data, with the mean and variance of the fitted Poisson distribution²² often taking a value in the region of 1.5 (meaning that the “central tendency” in many games is for the bulk of players to operate as if at S1 or else at S2).

To illustrate how the model might work, consider again the asymmetric matching pennies game from table 7.3. To keep the example simple, suppose that in a particular sample of 200 participants—100 assigned to play Row, 100 to play Col—there are four levels of players, distributed in the same way within each role: twenty S0 players; forty S1 players; thirty S2 players; and ten S3 players.

S0 players choose at random: so half of the twenty Row players choose U, the other half choose D; likewise, half of the twenty Col players choose L, half choose R.

S1 players suppose all other players are S0—which means that they expect each of their opponent’s strategies to be chosen with probability 0.5. On this basis, Col players are indifferent between L and R, so that,

²²The intuition suggested for using the Poisson distribution is as follows. Higher levels of reasoning require more intelligence/time/effort/working memory, so that out of any set of players who can reason at least at level k , the percentage who can also reason at level $k + 1$ falls as k rises—this is well captured by Poisson distributions.

on average, twenty of the forty play L while the other twenty play R. Meanwhile, as all S1 Row players expect L and R to be played with equal frequency by S0 Col players, all forty S1 Row players will choose U.

S2 players suppose all other players are either S0 or S1 and it is assumed that they judge the relative numbers of S0 and S1 correctly: i.e., one-third S0 and two-thirds S1. In this example, the proportions of S0 and S1 among Col players are not important for S2 Row players, since both S0 and S1 Col players are equally likely to play L or R. On this basis, all thirty S2 Row players will opt to play U. And actually, the exact proportion of S0 and S1 among Row players is not crucial for S2 Col players; as long as there are at least *some* S1 Row players, all of whom will choose U, the probability of U being played is greater than 0.5, in which case all thirty S2 Col players will choose R.

Finally, consider the S3 players. Since the payoffs for L and R are the same, Col players are only concerned with the probabilities associated with U and D; and as long as there is at least one S1 or S2 player, the probability of U being played will be greater than 0.5, in which case all ten S3 Col players will choose R. Meanwhile, for S3 Row players, the proportions of Col players *might* matter: for U to be strictly preferred, S3 Row players must expect the probability of L to be greater than 0.2—which in turn requires S3 Row players to judge that at least 40% of Col players are either S0 or S1 (since both of these types play L and R with equal probability). In fact, since two-thirds of the non-S3 Col players are S0 or S1, and since S3 Row players are assumed to judge that proportion correctly, this condition is easily satisfied, and all ten S3 Row players choose U.

On this basis, what would the experimenter observe? Half of the S0 and S1 Col players opt for each strategy, while all S2 and S3 Col players choose R; the overall pattern is therefore that L is played by thirty players and R by seventy players. Meanwhile, half of the S0 Row players opt for D but all the rest choose U, so that D is played by ten players and U by ninety players.

Of course, the precise numbers in this example depend on the initial distribution of types, here chosen so that it broadly resembles a truncated Poisson distribution, but rounded for simplicity of exposition. Simplified though it is, it serves to illustrate a number of points.

First, this CH approach can accommodate substantial departures from the Nash equilibrium prediction—and here in the direction observed by Ochs (1995) and also accommodated by QRE: namely, greater frequency of U and L than the 0.5 and 0.2, respectively, given by the mixed strategy equilibrium.

However, there is a sense in which such a CH model is not a thoroughgoing stochastic model. Apart from S0 players, who pick a strategy at random, all other types may be supposed to behave deterministically, acting in order to maximize expected payoffs on the basis of their beliefs about the other players.²³ These beliefs are a hybrid of the correct and the mistaken: players at each level are assumed to judge correctly the relative frequencies of all players at lower levels than themselves; but since they wrongly assume that there are no other players at the same or higher levels than themselves, they are liable to misjudge the actual probability with which any strategy is played. Thus, play is replete with errors but not, except in the case of S0 players, of the kind modeled by QRE or RP.

The example also shows that the CH model—in contrast with both Nash and QRE—is liable to be insensitive to changes in payoffs that do not alter the optimal strategy for any particular type of player. Suppose the payoff to Row from (U, L) was 9 rather than 4. The Nash mixed strategy equilibrium (somewhat counterintuitively) entails no change in the probability that Row plays U, but requires Col to reduce the probability of playing L to 0.1. Intuition, QRE, and the evidence (see, for example, McKelvey et al. 2000; Goeree et al. 2003) all suggest that Row will be more likely to play U as that payoff increases. But in the example above (and for many near-Poisson distributions with a mean in the same vicinity) there is no such implication in the CH model. Indeed, in this example, that payoff would have to fall below 2 before S3 Row players would switch to D.²⁴

So the CH approach has a number of appealing features—not least that it taps into the intuitively plausible idea that there are limits to the extent to which people are able (or at least willing) to think through the many levels of reasoning that some games might entail. On the other hand, in its current form—which might be characterized as a deterministic superstructure resting on a stochastic base—it has certain counterintuitive features and implications that are contrary to the evidence. But might there be scope for developing the model to allow for noise and impreci-

²³ This statement is arguably too sweeping: a generalization of CH, as we have presented it, might relax the assumption of deterministic play by higher-level players to allow some uncertainty among players at level 2 or above about the exact mix of lower-level players.

²⁴ On the basis of the figures in the example, S3 Row players believe that all S2 Col players will choose R but that half of all S0 and S1 Col players will choose L. Since S3 Row players believe that two-thirds of Col players are either S0 or S1, the probability of L being played is one-third: S3 Row players will therefore be indifferent between U and D when the payoff in the {U, L} cell is 2 and will strictly prefer U for all payoffs greater than 2.

sion that might complement the basic idea in a manner compatible with the data from experimental games?

It is beyond the scope of this chapter to do more here than indicate one or two broad lines of possible development. The most obvious candidate, given our earlier discussion, may be some modification to the deterministic superstructure of the model. As it stands, the model assumes that S1 players operate deterministically, given their assumption about their opponent *certainly* being S0. In turn, S2 players operate deterministically on the basis that they accurately predict the relative frequencies of players doing fewer steps of thinking and on this basis work out without noise or imprecision which strategy is optimal. This is a strong assumption; especially in a one-shot game where a player has had no opportunity to learn about the type or types of other players, one might expect that judgments about the likely proportions of other types might be a source of noise in the choice of strategy.

Something similar might be said for one further assumption: namely, that players sophisticated enough to do all that S2 or S3 players are supposed to do nevertheless disregard the possibility that any players operate at the same or higher levels as they themselves do. Camerer et al. (2004a) defend this assumption by appealing to some of the literature about "overconfidence" in judgment. But, against that, there is the possibility that players can conceive of some other players being at least as proficient at reasoning as they are. Indeed, much of the success displayed by participants in coordination games rests on the capacity of players to judge what others will choose when both players are required to play the same strategy. However, having players try to take account of the frequency of others with the same or higher powers of reasoning would present considerable analytical difficulties within a deterministic framework and for that reason this possibility is ruled out by assumption in existing models. Yet there may be scope for incorporating the possibility if uncertainties and variability in players' judgements about the probabilities with which their opponents will play the strategies available to them are considered.

7.3.4 Summary Remarks

In the previous three subsections we have seen two rather different ways of responding to the challenge of accommodating experimental data that appear to depart systematically from the predictions of standard game theory.

The first way is to take the standard theory but reformulate it in some stochastic form. As discussed in section 7.3.1, QRE does this by adding

an essentially Fechnerian error term to try to account for the observed departures—thereby achieving some improvement in fit, but relying on an equilibrium concept that entails substantial violation of stochastic dominance in order to get this improvement. Section 7.3.2 showed how an RP specification might share some of the broad characteristics of QRE but would not allow the same violations of stochastic dominance—thereby suggesting that if RP were the appropriate way of modeling noise and imprecision, the conventional core could *not* be reconciled with the experimental data. These conclusions run parallel with those drawn in section 7.2, where it was argued that although the Fechner model might appear to be able to reconcile *some* anomalies with the core provided by conventional theory, no such reconciliation was possible if the RP specification were used.

The second way of accommodating the experimental data is to modify various of the core assumptions. One version of this approach involves allowing for interactions between payoffs and modifying the nominal form of those payoffs to take account of comparative and/or interpersonal considerations. This may be seen as analogous to the various alternative individual decision theories such as regret theory (Bell 1982; Loomes and Sugden 1982, 1987) and disappointment theory (Bell 1985; Loomes and Sugden 1986) that allowed interactions between consequences as a means of accommodating violations of Independence and Transitivity. A rather different kind of modification of core assumptions is to take payoffs more or less as stated but to suppose that players' reasoning is bounded to different degrees—which might be regarded as broadly analogous to the explanations of individual decision anomalies in terms of simplifying heuristics. As things currently stand, the different ways of modifying various core assumptions do not build in noise and imprecision (except, in the case of CH, for S0 players). Considering how to specify such models stochastically will be a necessary part of testing them.

These points suggest that the construction and development of stochastic game-theoretic models, whether based on core game-theoretic assumptions or not, remains a crucial avenue for further research. Such research may run in parallel with, and draw on, research on the stochastic component of individual decision making. However, as games will also involve additional complexities, we return to the case of individual decision making in order to make some further points about possible forms of research on this topic.

7.4 Exploring Different Stochastic Specifications

The discussion in section 7.2 raised serious doubts about the appropriateness of a Fechnerian model of the stochastic component in individual decision making under risk and uncertainty. The suggestion made there was that, of the models currently available in this area, the RP model is a rather stronger candidate. However, that would imply that EU is the wrong core; and that in turn poses the question of how to apply RP to alternative core theories.

For example, consider how to apply RP to (some member of) the family of rank-dependent expected utility models. Such models usually entail some function $v(\cdot)$ that maps payoffs to subjective values and another function $\pi(\cdot)$ that transforms probabilities into decision weights according to a procedure that ensures FOSD.²⁵ The question then is, How much freedom should we allow to the set of value functions $v(\cdot)$ and to the set of probability weighting functions $\pi(\cdot)$? Taking cumulative prospect theory (Tversky and Kahneman 1992), for instance, the restrictions placed by the theory on the shape of $v(\cdot)$ are quite modest: the function is expected to be concave in the domain of gains, convex in the domain of losses, and steeper for losses than for corresponding gains, with a kink at zero. An RP form of such a core may require us to consider how we model the distributions of $v(\cdot)$ and $\pi(\cdot)$ and whether we require/disallow any particular relationships between the two. To date, so little attention has been paid to the question of the different candidate stochastic specifications that such questions take us into largely uncharted territory.²⁶

Moreover, when we move into the arena of experimental games, the issues appear to be even more wide open. As discussed in section 7.3, there are at least two plausible sources of noise: first, the same kind of noise manifested in individual decision experiments, relating to the translation of money payoffs into subjective utilities/values—this being even further complicated by possible considerations of interpersonal utility interactions not present in “games against nature”; second, the uncertainty about what other “types” of player one might interact with,

²⁵ An early form of this type of model was Quiggin’s (1982) “anticipated utility” theory. Subsequently, Starmer and Sugden (1989) proposed a form that incorporated a reference point and allowed losses to be treated differently from gains. Essentially the same idea is at the heart of Tversky and Kahneman’s (1992) cumulative prospect theory. Numerous other variants have also been proposed.

²⁶ A recent foray into part of this territory is reported by Stott (2006), who considers a number of combinations of different “core” specifications with various error models. However, he explicitly avoids the RP approach, confining attention to models that take an essentially Fechnerian approach.

and what the likelihoods are of encountering any particular type. Given the multiplicity of ways in which one could “not unreasonably” model how players translate the payoffs into utilities and how they could imagine other players to be likely to think and act, one might ask whether there are *any* patterns of behavior for which it would be impossible to construct a “not unreasonable” account.

One possible response to this is to recall the Lakatosian distinction, discussed in chapter 3, between progressive and degenerating research programs. On this account, merely mopping up anomalies with “not unreasonable” adjustments to protective belt assumptions is insufficient for progress; novel predictions and expanding empirical success are also required.

An alternative response is to try to investigate “noise” directly. How it operates in individual and/or interactive decision environments is essentially an empirical question. But the existing stock of evidence designed to address this issue is limited. Thus, a pressing task is to gather data about the nature and structure of the stochastic component of behavior. The question then—and this is the question that the rest of this section will address, albeit in an indicative rather than a comprehensive manner—is how that might be done.

7.4.1 Other Potentially Useful Forms of Data

The bulk of the existing relevant evidence from individual decision experiments takes the form of repeated pairwise choice data. As described earlier, such experiments have usually involved asking respondents to make large numbers of choices between pairs of lotteries, with particular pairs being presented on more than one occasion and with no feedback about the resolution of any choices until all decisions have been made. Then, typically, one pair will be selected at random from all those that have been presented and the respondent will be paid according to her choice in that case.

Arguably, this is about as close as one can get to taking random sample observations of a respondent’s preferences. Even so, as noted in section 7.2, there is some evidence of systematic change in the course of a session, with choices tending to become somewhat more risk averse in later responses. Quite what causes this is not known, but the patterns suggest that experience/familiarity has some impact even when there is only rather minimal feedback.

However, one drawback with discrete choice data of this kind is that they provide only rather limited information. All we observe on each occasion is which alternative was chosen, and the only visible “action”

occurs in cases where the choice changes from one occasion to another. In most cases, when no switch of choice occurs, we gain very little insight; and for any one individual it would require a great many repeated choices targeted in the vicinity of her switch-points to build up a picture of how the noise/error component of her preferences is configured.

It might seem that a much more efficient way of eliciting the information would be by means of a series of “equivalence” tasks in one of two forms: certainty equivalence and probability equivalence. For any lottery, the certainty equivalence task requires the respondent to state the sure sum of money she regards as exactly equally preferable to that lottery. Probability equivalence tasks require respondents to construct some other lottery that she regards as exactly equally preferable to the lottery being evaluated. One simple way of formulating this task is in terms of a “standard” or “reference” lottery with two payoffs, one of which is zero while the other is a sum greater than any of the payoffs in the lottery being evaluated.

Probability equivalence tasks have not been much used in economics experiments,²⁷ but there are many experiments that have involved certainty equivalents. Unfortunately, most of these do not provide data that can be used for our present purpose. Many individual decision experiments only ask respondents for a single valuation of each lottery. And although there are also some “repeated market” experiments that ask respondents to value the same lottery on a number of occasions in successive trading rounds, these generally provide considerable feedback between each round—feedback that is liable to exert substantial systematic influences upon later responses. Thus even though the repeated discrete choice data are not completely “pure,” the valuation datasets from repeated market experiments are even less like sets of random sample observations of respondents’ preferences.²⁸

In order to obtain equivalence data comparable to the kind of repeated choice data discussed earlier, it would be necessary to ask respondents to state a number of equivalences, with particular lotteries presented

²⁷ One area in which they *have* been more widely used is in the context of health state elicitation: in the simplest form of “standard gamble,” respondents are asked to set the probabilities of two outcomes of a health treatment—full health if the treatment succeeds and death if the treatment fails—such that they are indifferent between that risky prospect and the certainty of some intermediate health state. For further details, see Gafni (2005).

²⁸ Even when lotteries are not actually played out at the end of each round of trading, information about market prices (and sometimes about the distribution of bids and asks of other traders) is usually available to respondents and is liable to “shape” their subsequent responses. In addition, early bids/asks may be “contaminated” by strategic behavior, which may be modified to some extent by experience during subsequent rounds. For some evidence and discussion, see Loomes et al. (2003).

to them on more than one occasion within each session, and play out a randomly selected task only at the end of the session. However, we know of no substantial datasets that satisfy these requirements. In the absence of such datasets, we consider some data from an experiment that does not exactly fit this design but that may nevertheless provide some insights—as well as indicating some of the potential problems and pitfalls involved in collecting and analyzing equivalence data.

In a recent paper, Butler and Loomes (2007)—henceforth B&L—explored the role of imprecision in people’s preferences as a possible (part of the) explanation of the preference reversal phenomenon (see box 4.3 (p. 156)). Their design revolved around two lotteries: a *P*-bet offering a 0.70 chance of \$24 (and a 0.30 chance of zero), and a *\$*-bet offering a 0.25 chance of \$80 (and a 0.75 chance of zero).²⁹ The experiment sought, among other things, to elicit from each respondent a certainty equivalent and a probability equivalent for each lottery. The elicitation of those equivalences was implemented as follows. The bet in question—say, the *\$*-bet—was fixed and labeled A, while the other option (a sure sum of money in the case of the certainty equivalence task, or some chance of a \$160 payoff in the case of the probability equivalence task) was labeled B and was initially set at some “extreme” value.³⁰ Both alternatives were displayed on a computer screen, and the individual was then asked to respond in one of four ways: if they “definitely preferred” option A, this was coded as a 1; if they “probably preferred” A, this was recorded as a 2; “probably preferring” B was recorded as a 3; and a “definite preference” for B was coded as a 4. Then the value in B was altered—progressively increasing if it initially started at the bottom extreme, progressively falling if it initially started high. Thus the typical respondent began by being sure they preferred one alternative and ended up being sure they preferred the other, and in half of the cases this entailed a transition from A to B while in the other half the movement was from B to A. In this way, each respondent not only effectively stated a certainty and probability equivalent for the target bet (the point at which they switched from probably preferring one to probably preferring the other, that is, the 2 ↔ 3 switch-point), but also indicated the interval (between 1 ↔ 2 and 3 ↔ 4) over which they considered themselves to be less than sure about their equivalence. We shall refer to these latter intervals as “imprecision intervals.”

²⁹These payoffs were in Australian dollars.

³⁰In the certainty equivalence task for the *\$*-bet, for example, for half the sample the alternative was initially set as a sure payoff of \$1, while for the other half it started at \$80; in the probability equivalence task, the initial alternative was either a 0.01 chance of \$160 or else a 0.25 chance of \$160.

Someone who was totally sure of their preference could go straight from recording a definite preference for one option to a definite preference for the other, in which case indifference for them would be represented by the 1 \leftrightarrow 4 switch-point. In the event, however, all but a handful of respondents recorded at least some 2s and also some 3s in each exercise. This is consistent with most people not only exhibiting variability in their decisions but also being aware of the possibility that they might reach different decisions on different occasions. And although it is not known exactly what individuals were thinking when they changed from recording a “definite” preference to a “probable” one, and vice versa, these responses may still provide useful data for evaluating competing ways of modeling the stochastic component in decision behavior.

The main patterns that emerged were as follows. In the case of the certainty equivalence task, the mean imprecision interval for the *P*-bet was less than one-third of the corresponding interval for the *S*-bet (approximately \$6 compared with \$20). In the probability equivalence task, the mirror-image pattern was observed: for the *S*-bet, the average interval between the 1 \leftrightarrow 2 and 3 \leftrightarrow 4 switching points was roughly 0.07, compared with a figure of 0.20 for the *P*-bet.³¹

Could these results be reconciled with a Fechnerian specification of EU—even one that allows the variance of ε to be some function of the characteristics of a lottery and therefore to vary from one lottery to another? To see that the answer is probably “No,” consider the following.

One way of thinking about imprecision intervals in Fechnerian terms is as confidence intervals. Under Fechnerian assumptions, in any choice between two lotteries there is some probability that the individual will choose the alternative whose “true EU” is lower, with this probability becoming smaller as the difference between the true EUs of the two lotteries becomes larger. It seems reasonable to suppose that individuals’ expressions of confidence in their choice should become stronger as the chance of picking the “wrong” alternative becomes smaller, and so we might think of the statement that they “definitely prefer” one lottery as indicating that they consider the chance of being mistaken lying below some threshold, while “probably preferring” might signify that they think the chance of being mistaken is greater than that threshold. It would be in keeping with the spirit of Fechner to suppose that a given

³¹ Those aggregate patterns were also evident at the level of the individual: for seventy-three of the eighty-nine individuals, the imprecision intervals for their certainty equivalents of the *S*-bet were strictly greater than the corresponding intervals for the *P*-bet; while in the probability equivalence task, seventy-one of the eighty-nine exhibited strictly wider intervals for the *P*-bet than for the *S*-bet.

individual applies the same threshold to all choices.³² Thus an imprecision interval can be thought of, in Fechnerian terms, as an individual's confidence interval over the joint distribution of the noise associated with a choice between any two alternatives.

Now consider the case of an EU-maximizing individual who is truly indifferent between two lotteries, F and G: the same true EU therefore maps to the same true certainty, denoted by C^* , and the same true probability equivalent, denoted by p^* (where this is the probability of receiving the high payoff in a reference lottery R, with that payoff high enough to guarantee $p^* < 1$). In addition, suppose that for some reason the variance of ε is greater for F than for G (with both distributions symmetrical around the same true EU).

Consider first the B&L method of eliciting probability equivalences for the two lotteries, starting with (say) a high value of p that is progressively reduced from a level where the individual definitely prefers R to a level where she definitely prefers either F or G. At every level of p , the joint distribution of the error terms in the {R, F} choice will be higher than the corresponding joint distribution for {R, G}, with the result that for any individual applying the same threshold "level of significance" to both, the confidence interval/imprecision interval will be wider for F than for G.

Now consider the B&L method of eliciting certainty equivalences, starting with (say) a sure sum large enough for it to be definitely preferred to the lottery and then progressively reduced until it is low enough for the lottery to be definitely preferred. Whether or not we assume some noise associated with the perceived utility of the sure sum,³³ at any level of C the joint distribution of the error terms in the {C, F} choice will be higher than the corresponding joint distribution for {C, G}. Thus, just as with the probability equivalent elicitation, each individual's confidence interval/imprecision interval expressed in certainty equivalent terms will be wider for F than for G.

In short, under the assumptions of the Fechner model, if the ε associated with lottery F has greater variance than that for the ε associated with lottery G, this would tend to be reflected in broader confidence/imprecision intervals, whether measured via the probability of some reference payoff or via sure sums of money.

So if the intervals identified by B&L are proxies for such confidence intervals, the patterns in the experimental data are incompatible with

³²We can, of course, allow that the degree of confidence separating a "definite" from a "probable" preference might vary from one individual to another.

³³In the case of Blavatsky's (2007) variant, the utilities of sure sums are always perceived without any error at all. But this assumption is not necessary to derive the results presented here.

the Fechner model: instead of one lottery being associated with a wider interval whichever way the interval is measured, the average imprecision interval for the \$-bet was more than three times wider than the interval for the P -bet when elicited by certainty equivalence, while the \$-bet interval was barely one-third of the width of the P -bet interval when these were elicited via the probability equivalence procedure.

By contrast, the ways in which the relative sizes of intervals varied in that experiment is much more easily accommodated by the RP framework. To illustrate this, consider an example of a risk-averse EU maximizer whose set of $u(\cdot)$ functions all take the form $u(x) = x^\beta$, where β is distributed over some range. Suppose that a particular individual acts as if sure that her β is not less than 0.6 and not more than 0.8, but reports herself as only having a “probable” preference in cases that fall within that range. Applying those two values of β to give the ends of her imprecision intervals, her interval for the probability equivalent of B&L’s P -bet would lie between 0.154 and 0.224—that is, an interval of 0.070—while her interval for the \$-bet would lie between 0.144 and 0.165—that is, an interval of 0.021, which is less than one-third of the P -bet interval. Applying those same two “end” values of β to the certainty equivalence task would give an interval of \$2.12 for the P -bet and \$6.21 for the \$-bet, so that in this case the \$-bet interval is almost three times as wide as its P -bet counterpart. Stylized though this example may be, it serves to show how the pattern of the imprecision intervals reported by B&L is much more easily reconciled with the RP model than with any reasonably straightforward version of the Fechner specification.³⁴

Although it would be rash to draw strong conclusions from a single exploratory study, the approach developed by B&L may provide insights into the imprecise and/or stochastic nature of people’s preferences—in which case, extensions of the approach might enable us to build a more complete picture of the nature of noise/imprecision in the behavior of experimental subjects.

On the other hand, the type of procedure used by B&L is not uncontroversial, for (at least) two reasons. First, it relies upon introspection about confidence in a decision, which is a difficult notion to pin down and interpret. Second, there is no obvious way of linking the $1 \leftrightarrow 2$ and $3 \leftrightarrow 4$ switch-points to standard monetary incentives. As chapter 6 explained,

³⁴In the light of what was said earlier about the rejection of EU as the core theory, conditional on RP being the appropriate model of the stochastic component, the use of EU in the example here is not meant to suggest that EU can be rehabilitated; rather, an EU core is deployed to keep things simple and make it clear that the result does not depend on invoking some more complicated core theory involving nonlinear probability transformations and/or more complex value functions.

we do not take the view that information obtained via a procedure that cannot be made incentive compatible in the eyes of orthodox theory should necessarily be ruled inadmissible. But there is a strong presumption among some experimental economists in favor of incentives. In view of this, a possible task for future research is to explore whether the possibilities suggested by the B&L procedure can be substantiated by experimental designs that *can* employ more standard incentive-compatible mechanisms.

7.5 Concluding Remarks

Even when making choices between pairs of the most basic and well-defined lotteries, many participants in experiments display variability that may be modeled as stochastic. If this is true for individual choices under risk, it is likely to be at least as significant in the behavior observed in experimental games where there is additional uncertainty and room for variable judgment about the likelihood of different strategies being played.

To date, this issue has received less attention than we believe it deserves. Often, statistical tests are applied that implicitly assume some error structure, but these are taken “off the shelf,” perhaps with too little thought about the appropriateness of the assumptions that underlie them. And to the extent that the issue has been more explicitly addressed, the usual approach has been to model noise as an “add-on” in the Fechnerian tradition.

We have argued that such specifications, while they may seem quite “natural” to economists and econometricians, are not neutral or uncontroversial. Indeed, there are grounds for thinking that, at least in the context of individual choice and experimental games, the Fechner specification is conceptually flawed and empirically inadequate. In particular, such a specification is liable to (greatly) overestimate the frequency of violations of *transparent* dominance (which are actually quite rare in individual risky choice experiments) and, in conjunction with standard core theories, entails a number of patterns that do not cohere with the evidence.

To demonstrate that the Fechnerian approach is not the only possible way to model the stochastic component in people’s preferences, we have discussed the potential of the RP approach as an alternative. Whether RP can provide an entirely satisfactory—or even a less unsatisfactory—account is still open to debate. But what the comparisons between RP

and Fechner have shown beyond doubt is that different stochastic specifications of the same core theory can produce radically different implications, and the message is clear: the choice of stochastic specification is not an “optional extra” but is absolutely central to the analysis and interpretation of experimental data in key areas of economic behavior and theory testing.

While this chapter has given most prominence to the Fechner and RP approaches, we have not intended to suggest that they are the only ways of modeling the stochastic nature of preferences. Other formulations—for example, Busemeyer and Townsend’s (1993) “decision field theory”—may offer additional and perhaps rather different insights. The main purpose of the chapter has been to draw attention to how little has yet been done in this important area of enquiry and to stimulate further debate and research. While we might stop a *little* short of Chairman Mao’s invocation to “let a hundred flowers bloom; let a hundred schools of thought contend,” it is right in spirit. There is an important symbiotic relationship to be explored here: experimental economists must understand the stochastic elements of behavior better if we are to analyze and interpret our data appropriately; and yet it is hard to see how to advance our understanding of those elements without further experimentation, including the development and use of methods that may reach beyond some of the orthodoxies that we have examined in the course of this book.