European Review of Agricultural Economics Vol **38 (3)** (2011) pp. 335–360
doi:10.1093/erae/jbr029
Advance Access Publication 15 July 2011

Experimental methods and the welfare evaluation of policy lotteries

Glenn W. Harrison*

Georgia State University, Atlanta, USA

Received April 2011; final version accepted May 2011

Review coordinated by Thomas Heckelei

Abstract

Policies impose lotteries of outcomes on individuals, since we never know exactly what the effects of the policy will be. In order to evaluate alternative policies, we need to make assumptions about individual preferences, even before social welfare functions are applied. There are two broad ways in which experimental methods are used to evaluate policy. One is to use experiments to estimate individual preferences, valuations and beliefs and use those estimates as priors in policy evaluation. The other is to use randomisation to infer the effects of policy. The strengths, weaknesses and complementarities of these approaches are reviewed.

Keywords: experimental methods, risk attitudes, subjective beliefs, policy evaluation

JEL classification: D03, D04, D81

1. Introduction

Policies impose lotteries of outcomes on individuals, since we never know exactly what the effects of the policy will be. In order to evaluate alternative policies, we therefore need to make some assumptions about individual preferences, even before social welfare functions are applied. One simply cannot make claims that individual welfare is improved unless one knows what risk attitudes, discount rates and subjective beliefs drive behaviour. And then one has to untangle descriptive characterisations from normative characterisations. Should the subjective judgements of 'experts' be substituted for those of the individuals ultimately affected by the policy? Or the risk attitudes and discount rates of 'society'? These are important value judgements, but before they are to be adopted we have to know how much of a difference they make compared with the 'own tastes and beliefs' of the parties affected.

*Corresponding author: Department of Risk Management and Insurance, Robinson College of Business, Georgia State University, 35 Broad Street NW, Atlanta, GA 30303. E-mail: gharrison@gsu.edu

Consider the humble question of the welfare valuation of some new insurance product, such as the 'micro-insurance' products being offered in developing countries. In general these policies are evaluated by the metric of product take-up. About the only virtue of this metric is that it is easy to measure. An insurance product involves the individual giving up a certain amount of money *ex ante* some event in the expectation of being given some money in the future if something unfortunate occurs. Welfare evaluation requires that one knows risk and time preferences of the individual, since the benefits of the product are risky, and in the future, while the costs are normally certain and up front. We must also know the subjective beliefs that the individual used to evaluate the product, and let us not even start to assume any uncertainty aversion or ambiguity aversion. Of course, there is a 'revealed preference' argument that if the product is (not) taken up it was perceived to be a positive (negative) net benefit. But that is only the starting point of any serious welfare evaluation. What if the subjective beliefs were off, in the sense that the individual would revise them if given certain information?

Instead of making *a priori* assumptions about those preferences that are likely to be wrong, there are two broad ways in which experimental methods can be used to evaluate policy. One is to use experiments to estimate individual preferences, valuations and beliefs, and use those estimates as priors in the evaluation of policy. The other approach is to undertake deliberate randomisation, or exploit accidental or natural randomisation, to infer the effects of policy. The strengths and weaknesses of these approaches are reviewed, and their complementarities identified.

2. The concept of a policy lottery

The place in which the concept of a policy lottery appears in its most explicit form is in the use of computable general equilibrium (CGE) models to evaluate the effects of public policy. Those policies range over domestic tax reforms, agricultural policy reforms pursuant to global trade agreements, unilateral trade policies and unilateral and multilateral carbon tax reforms. One of the hallmarks of these CGE models was an explicit recognition that many of the structural parameters of those models were uncertain, and that policy recommendations that came from them amounted to a policy lottery in which probabilities could be attached to a range of possible outcomes. Recognition that the simulated effects of policy on households were uncertain, because the specific parameters of the model were uncertain, meant that a proper welfare analysis needed to account for the risk attitudes of those households.

Related to this dimension of these simulated results, in many cases there were nontrivial intertemporal tradeoffs: foregone welfare in the short-term in return for longer term gains. Indeed, this tradeoff is a common feature of dynamic CGE policy models (e.g. Harrison *et al.*, 2000). Obviously the proper welfare evaluation needed to also account for the subjective discount rates that those households employed. For example, one of the policy issues of interest to the Danish government was why Danes appeared to 'underinvest' in higher education (see Lau, 2000).

A policy lottery is a representation of the predicted effects of a policy in which the uncertainty of the simulated impact is explicitly presented to the policy maker. Thus when the policy maker decides that one policy option is better than another, the uncertainty in the estimate of the impact has been taken into account. This is uncertainty in the *estimate of the impact*, and not necessarily uncertainty in the *impact itself*. But in the limited information world of practical policy-making such uncertainties are rife.[1]

We first illustrate in an explicit, structural manner the nature of the policy lottery we have in mind, using a tax policy setting and a climate policy setting from the *Stern Report* on climate change. Although these are explicit simulation models that one would not expect to see in most policy settings, they illustrate clearly the type of information one needs to make an informed decision. Even if informed, having lots of explicit structure does not of course make the decision the right one. Nor does one need such structure to see the point that one has to worry about modelling uncertainty in the formation of policy.

Or does one? Is it possible to arrive at 'evidence based' policy conclusions without structure? Some have argued that it is indeed possible, using Randomized Control Trials (RCT) of policy interventions. We consider the strengths and weaknesses of this approach below, but the two examples of policy lotteries are deliberately chosen to involve interventions of some policy significance that could not be studied using an RCT.

2.1. A detailed example of a tax policy lottery

We illustrate the concept of a policy lottery using the CGE model documented in Harrison *et al.* (2002a). This static model of the Danish economy is calibrated to data from 1992. The version we use has 27 production sectors, each employing intermediate inputs and primary factors to produce output for domestic and overseas consumption. A government agent raises taxes and pays subsidies in a revenue-neutral manner, and the focus of our policy simulation is on the indirect taxes levied by the Danish government.[2] A representative government household consumes goods reflecting public expenditure patterns in 1992. The simulated policy effects are different across several private household types. The model is calibrated to a wide array of empirical and *a priori* estimates of elasticities of substitution using nested constant elasticity of substitution specifications for production and utility functions.

[1] For example, see Desvousges, Johnson and Banzhaf (1999). The limitation on information can derive from the inherent difficulty of modelling behavioural or physical relationships, from the short-time frame over which the model has to be developed and applied, or both.

[2] Revenue neutrality is defined in terms of real government revenue, and does not imply welfare neutrality.

The model represents several different private households, based on the breakdown provided by Statistics Denmark from the national household expenditure survey. For our purposes, these households are differentiated by family type into seven households: singles younger than 45 without children, singles older than 45 without children, households younger than 45 without children, households older than 45 without children, singles with children, households with children and where the oldest child is 6 or under and households with children and where the oldest child is between 7 and 17. The model generates the welfare impact on each of these households measured in terms of the equivalent variation in annual income for that household. That is, it calculates the amount of income the household would deem to be equivalent to the policy change, which entails changes in factor prices, commodity prices and expenditure patterns. Thus the policy impact is some number of Danish kroner, which represents the welfare gain to the household in income terms.

This welfare gain can be viewed directly as the 'prize' in a policy lottery. Since there is some uncertainty about the many parameters used to calibrate realistic simulation models of this kind, there is some uncertainty about the calculation of the welfare impact. If we perturb one or more of the elasticities, for example, the welfare gain might well be above or below the baseline computation. Using randomised factorial designs for such sensitivity analyses, we can undertake a large number of these perturbations and assign a probability weight to each one (Harrison and Vinod, 1992). Each simulation involves a random draw for each elasticity, but where the value drawn reflects estimates of the empirical distribution of the elasticity.[3] We undertake 1,000 simulations with randomly generated elasticity perturbations, so it is as if the household faces a policy lottery consisting of 1,000 distinct prizes that occur with equal probability 0.001. The prizes, again, are the welfare gains that the model solves for in each such simulation.

Figure 1 illustrates the type of policy lottery that can arise. In this case we consider a policy of making all indirect taxes in Denmark uniform, and at a uniform value that just maintains the real value of government expenditure. Thus we solve for a revenue-neutral reform in which the indirect tax distortions arising from inter-sectoral variation in those taxes are reduced to zero. Each box in Figure 1 represents 1,000 welfare evaluations of the model for each household type. The large dot is the median welfare impact, the rectangle is the inter-quartile range and the whiskers represent the range of observed values. Thus we see that the policy represents a lottery for each household, with some uncertainty about the impacts.

If a policy-maker were to evaluate the expected utility to each household from this policy, he would have to take into account the uncertainty of the estimated outcome and the risk attitudes of the household. The traditional

[3] For example, if the empirical distribution of the elasticity of substitution is specified to be normal with mean 1.3 and standard deviation 0.4, 95 per cent of the random draws will be within $\pm 1.96 \times 0.4$ of the mean. Thus one would rarely see this elasticity take on values greater than 3 or 4 in the course of these random draws.

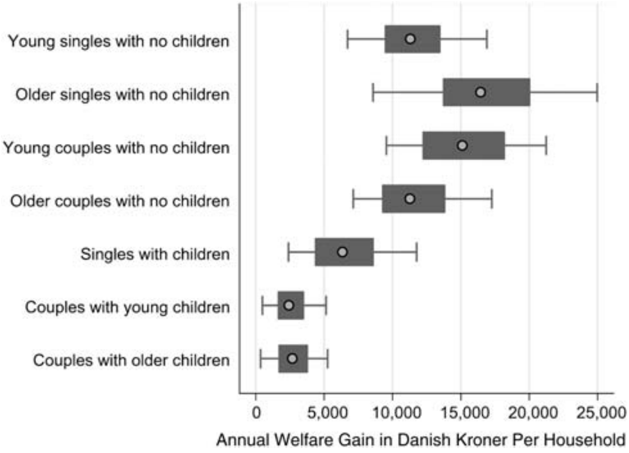

Fig. 1 An illustrative policy lottery.

approach in policy analysis is to implicitly assume that households are all risk-neutral and simply report the average welfare impact. But we know from experimental results reported in Harrison, Lau and Rutström (2007) that these households are not risk neutral. Assume a constant relative risk aversion (CRRA) utility specification for each household. We can stratify the raw elicited CRRA intervals according to these seven households and obtain CRRA estimates of 1.17, 0.48, 0.79, 0.69, 0.76, 0.81 and 0.95, respectively, for each of these households. In each case these are statistically significantly different from risk neutrality.

Using these CRRA risk attitude estimates, it is a simple matter to evaluate the utility of the welfare gain in each simulation, to then calculate the expected utility of the proposed policy and to finally calculate the certainty-equivalent welfare gain. Doing so reduces the welfare gain relative to the risk-neutral case, of course, since there is some uncertainty about the impacts. For this illustrative policy, this model, these empirical distributions of elasticities, and these estimates of risk attitudes, we find that the neglect of risk aversion results in an overstatement of the welfare gains by 1.6, 1.4, 1.8, 1.1, 5.1, 4.6 and 7.9 per cent, respectively, for each of the households. Thus a policy maker would overstate the welfare gains from the policy if risk attitudes were ignored.

Tax uniformity is a useful pedagogic example, and a staple in public economics, but one that generates relatively precise estimates of welfare gains in most simulation models of this kind. It is easy to consider alternative realistic policy simulations that would generate much more variation in welfare gain, and hence larger corrections from using the household's risk attitude in policy evaluation. For example, assume instead that indirect taxes in this model were reduced across the board by 25 per cent, and that the government effected lump-sum side payments to each household to ensure that no household had less than a 1 per cent welfare gain. In this case, plausible elasticity configurations for the model exist that result in very large welfare gains for some households.[4] Ignoring the risk attitudes of the households would result in welfare gains being overstated by much more significant amounts, ranging from 18.9 to 42.7 per cent depending on the household.

These policy applications point to the payoff from estimating risk attitudes, as we do here, but they are only illustrative. A number of limiting assumptions obviously have to be imposed on our estimates for them to apply to the policy exercise. First, we have to assume that the estimates of CRRA obtained from our experimental tasks defined over the domain of prizes up to 4,500 DKK apply more widely, to the domain of welfare gains shown in Figure 1.[5] Given the evidence from our estimation of the Expo-Power function, reported in Harrison, Lau and Rutström (2007), we are prepared to make that assumption for now. Obviously one would want to elicit risk attitudes over wider prize domains to be confident of this assumption, however. Second, we only aggregate households into seven different types, each of which is likely to contain households with widely varying characteristics on other dimensions than family types. Despite these limitations, these illustrations point out the importance of attending to the risk preference assumptions imposed in policy evaluations. Recent efforts in modelling multiple households in CGE have been driven by concerns about the impacts of trade reform on poverty in developing countries, since one can only examine those by identifying the poorest households: see Harrison *et al.* (2003) and Harrison, Rutherford, Tarr and Gurgel (2004). Clearly one would expect risk aversion to be a particularly important factor for households close to or below the absolute poverty line.

It might be apparent that we would have to conduct field experiments with a sample representative of the Danish population in order to calibrate a CGE model of the Danish economy to risk attitudes that were to be regarded as having any credibility with policy-makers. But perhaps this is not so obvious to academics, who are often happy to generalise from convenience samples.

2.2. Stern's climate change policy lotteries

The idea of a policy lottery plays a central role in the *Stern Review on the Economics of Climate Change* (Stern, 2007). It stresses (Stern, 2007: 163) the need to have a simulation model of the economic effects of climate change that can show stochastic impacts. In fact, any of the standard climate simulation models can easily be set up to do that, by simply undertaking a systematic sensitivity analysis of their results. The *Review* then proposes an 'expected utility analysis' of the costs of climate change (Stern, 2007: 173ff.) which is effectively the same as viewing climate change impacts as

[4] For example, if the elasticity of demand for a product with a large initial indirect tax is higher than the default elasticity, households can substitute towards that product more readily and enjoy a higher real income for any given factor income.

[5] At the time of the experiments, June 2003, DKK 7.43= EUR 1.

a lottery. When one then considers alternative policies to mitigate the risk of climate change, the 'expected utility analysis' is the same as our policy lottery concept, with the addition that the baseline business as usual (BAU) path is also stochastic.

Stern (2007: Chapter 6) describes a formal simulation model to estimate the cost of climate change. He extends existing simulation models that predict 2°C to 3°C of warming over the next century at a cost of 0 to 3 per cent of GDP by considering additional, low-probability states of the world resulting in 5°C to 6°C of warming. The estimated costs for his six states of the world, which are reported in balanced growth equivalents, range at their mean from a 2.1 to a 14.4 per cent loss in current consumption under a BAU baseline, which is one without regulatory price or quantity constraints placed by governments. Much of the rest of the *Stern Report* argues that this BAU loss in per capita consumption is a conservative estimate and would be much larger if direct non-market damages to human health and the environment, non-linear climate feedback or distributional impacts to poor nations were factored into the model.

The quantitative analysis of the *Stern Report* differs from earlier models in the manner that it calculates the monetary cost of climate change and in how it interprets those costs. It explicitly incorporates the stochastic element of climate change science by simulating costs across a wide range of possible outcomes, including those that are extremely low-probability and highly damaging. It also subtracts items from GDP such as air conditioning and flood defence that may actually increase as a result of rising temperatures, arguing that this method makes reported losses in GDP more accurate measures of income loss rather than output loss. The costs across all possible states of the world are then interpreted using expected utility theory (EUT). He finds the utility or social welfare for each state of the world and assigns a subjective probability of that state occurring. The EU value then becomes the weighted average of each utility value and it's corresponding subjective probability.

Stern (2007) presents simulation results for two climate scenarios and three categories of economic impact. The baseline climate scenario is designed to give results consistent with the Intergovernmental Panel on Climate Change, while the high climate scenario adds to this the risk of amplifying feedbacks in the climate system at higher temperatures. The high climate scenario assumes a higher probability of larger temperature change. For example, the baseline predicts mean warming of 3.9°C by 2100 relative to pre-industrial average temperature with a 90 per cent confidence interval of 2.4°C–5.8°C. The high climate scenario estimates a mean warming of 4.3°C by 2100 with a 90 per cent confidence interval of 2.6°C–6.5°C. The range of possible temperature results is larger and skewed upwards, as expected, in the high climate scenario.

3. The role of experiments

Experiments can help inform the evaluation of policy lotteries in two ways. The first is by providing some guidance as to latent structural parameters needed to complete the welfare evaluation. The second is by bypassing the need for all of this structure, in an agnostic manner, and 'letting the data speak for itself' with minimal theoretical assumptions.

It is worth identifying the various types of experiments in wide use. Harrison and List (2004) propose a taxonomy to help structure thinking about the many ways in which experiments differ. At one end of the spectrum are *thought experiments*, which can be viewed as the same as any other experiment but without the benefit of execution (Sorenson, 1992). Then there are conventional *laboratory experiments*, typically conducted with a convenience sample of college students and using abstract referents.[6] Then there are three types of field experiments. *Artefactual field experiments* are much like lab experiments, but conducted with subjects that are more representative of a field environment. *Framed field experiments* extend the design to include some field referent, in terms of the commodity, task or context. *Natural field experiments* occur without the subject knowing that they have been in an experiment. Then we have *social experiments*, where a government agency deliberately sets out to randomise some treatment. Finally, there are *natural experiments*, where some randomisation occurs without it being planned as such: serendipity observed. Randomisation can be used in every one of these types, and is more a method of conducting experiments rather than a defining characteristic of any one type of experiment in the field, as some have suggested. Nor are these categories intended to be hard and fast: one can easily imagine intermediate categories, such as the *virtual experiments* of Fiore *et al.* (2009), with the potential of generating both the internal validity of lab experiments and the external validity of field experiments.

3.1. Estimating preferences and beliefs

There are three fundamental, behavioural 'moving parts' in almost any decision of importance: risk attitudes, time preferences and subjective beliefs. Experimental economists now have a robust set of tools to elicit each of these, although controversies remain, as expected in foundational concepts such as these.

Risk attitudes refer to the risk premium that individuals place on lotteries. The familiar diminishing marginal utility explanation of EUT provides one characterisation of the risk premium, and allows a wide range of flexible utility functions to be estimated. But it is a simple matter to also allow for probability weighting to explain the risk premium: 'pessimistic' attitudes towards probabilities can just as easily account for risk aversion.[7] Similarly, it is possible to extend the estimation to allow for sign-dependent preferences,

[6] A referent is an object or idea to which a word or phrase refers.

[7] The logic is easy to see. Assume lotteries defined solely over gains, and a linear utility function just to remove the effect of diminishing marginal utility. Then if the weighted probability is always equal to or less than the actual (objective or subjective) probability, the EU based on

whereby ‘losses’ are evaluated differently than ‘gains’. We add quotation marks for losses and gains because the Achilles Heel of sign-dependent models is the specification of the reference point, and this is the subject of considerable debate. All of these approaches simply decompose and explain the risk premium in different ways, and build on the approach before it. Experimental and econometric methods for the estimation of risk attitudes using all of these approaches are relatively well-developed: see Harrison and Rutström (2008) for an extensive survey.

There is also considerable evidence that behaviour towards risky lotteries is not characterised by just one model of decision-making under risk. This evidence comes from mixture specifications that allow two or more models of decision-making under risk, and let the data determine the mixture probability of each mode. These specifications, in rich and poor countries, in the lab and the field, show a remarkable combination that is close to 50:50 of *both* EUT *and* non-EUT characterisations (e.g. Harrison and Rutström, 2009; Harrison, Humphrey and Verschoor, 2010). This finding is likely to vary from domain to domain, and population to population, but offers a much richer characterisation of behaviour than the usual approach favoured by economists.[8]

Recent extensions include attention to the problem of the presence of ‘background risk’ affecting decisions over foreground risk (e.g. Harrison, List and Towe, 2007). For example, it makes little sense to evaluate the value of a statistical life without worrying about the confound of compensating differentials for non-fatal injuries: what does not kill often injures. A further extension to multi-variate, or multi-attribute, risks promises greater insight into risk management over traded and non-traded assets in the individual’s portfolio (e.g. Andersen *et al.*, 2011b).

Time preferences are also now relatively well understood. The first generation of experiments used loose procedures by modern standards, often relying on the elicitation of present values using Fill-In-The-Blank (FIB) methods that have notoriously poor behavioural properties. This literature is characterised by the need to use scientific notation to summarise estimated astronomic discount rates, a sure sign that something was wrong with behaviour, experimental design or inferential methods. Frederick, Loewenstein and O’Donoghue (2002) summarise the literature up to this point. The second generation of experiments moved towards binary choice tasks to ensure incentive compatibility, albeit at the loss of information precision (if the FIB methods behaved the way theorists advertised them, which was not the case) and stakes that were more substantial. Inferred discount rates were now at the level of consumer credit cards: high, but believable (e.g. Coller and Williams, 1999; Harrison, Lau and Williams,

these weighted probabilities will be less than the EV based on the actual probabilities, hence there is a risk premium.

[8] In effect, the usual methodological approach is akin to running a horse race, declaring a winner, maybe by a nose and shooting all of the losing horses. The fact that one of these losers might have done better on a different, wetter track is ignored.

2002b). The third generation of experiments recognised that discount factors equalise time-dated utility, and not time-dated money, so one needed to account for diminishing marginal utility when inferring discount factors. This is a simple matter of theory, from the conceptual definition of a discount factor. Jensen's Inequality does the rest theoretically: inferred discount rates must be lower if one has a concave utility function than if one assumes a linear utility function. Appropriate experimental designs and econometric inferences then simply quantify this insight from theory, with a dramatic reduction in estimated discount rates down to 10 per cent or even lower (e.g. Andersen *et al.*, 2008).

Quite apart from the level of discount rates, there appears to be no support for 'hyperbolicky' specifications of the discounting function in field data (e.g. Andersen *et al.*, 2011a). This does not mean that exponential specifications are appropriate for all populations, just that the monolithic presumption in favour of non-exponential specifications is not supported by the data.

Subjective beliefs can be elicited using scoring rule procedures that have a venerable tradition, such as Savage (1971). These procedures do require that one corrects for risk attitudes, and only directly elicits true subjective beliefs under the assumption of risk neutrality. But it is a relatively simple matter to condition inferences about beliefs on the estimated risk attitudes of individuals, by combining experimental tasks that allow one to identify the risk attitudes independently of the task that elicits subjective beliefs (e.g. Andersen *et al.*, 2010, 2011c). One can also use generalisations of these scoring rules to elicit whole subjective probability distributions, rather than just one subjective probability (e.g. Mathieson and Winkler, 1976, for the theory). This area is the least developed of the three, but the experimental tools are in place for rigorous elicitation, and are being widely applied.

It should be stressed that there are also many loose claims about how one can elicit risk attitudes, time preferences and subjective beliefs 'on the cheap' with simpler methods. In some cases these are hypothetical survey methods, with no theoretical claim to be eliciting anything of interest. In other cases these are experimental methods that rely, as noted, on tasks that are simply not incentive compatible: subjects could exploit the experimenter, for gain, by deliberately misrepresenting their true preferences. Or experimenters use FIB elicitation methods that have known behavioural biases, as noted above. The fact that experimenters assert that these problems did not arise says nothing about whether they do. The existence of relatively transparent, incentive compatible methods leads one to wonder why one would risk using other methods.

It is appropriate that all of these methods were first developed in laboratory environments, and that the econometric procedures for estimation of preferences and beliefs first refined in that setting. Lab experiments give us control, if designed and executed correctly. If we cannot identify the conceptually correct measure in that setting, we cannot hope to do so in more complicated field settings. But there is a relatively easy bridge between the lab and the field, as stressed by Harrison and List (2004), so that both are complementary ways to make inferences (Harrison, Law and Ruström, 2011).

3.2. Letting the data speak for itself

Randomised evaluations, inspired by the RCTs literature in health, have become popular in economics. They involve the deliberate use of a randomising device to assign subjects to treatment, or the exploitation of naturally occurring randomising devices. Good reviews of the methodology are contained in Duflo (2006), Duflo and Kremer (2005), Duflo, Glennerster and Kremer (2007) and Banerjee and Duflo (2009). Complementary econometric strategies are well described in Angrist and Pischke (2009).

One of the claimed advantages of randomisation is that the evaluation of policies can be ‘hands off’, in the sense that there is less need for maintained structural assumptions from economic theory or econometrics. In many respects this is true, and randomisation does indeed deliver, on a good, asymptotic randomising day, orthogonal instruments to measure the effect of treatment. This has been well known for a long time in statistics, and of course in the economics experiments conducted in laboratories for decades. But it is apparent that the case for randomisation has been dramatically oversold: even if the original statements of the case have the right nuances, the second generation of practitioners seems to gloss those. Words such as ‘evidence based’ ‘assumption free’ are just marketing slogans, and should be discarded as such. Excellent critiques by Rosenzweig and Wolpin (2000), Keane (2010), Leamer (2010), Heckman (2010), Deaton (2010), and spirited defences by Imbens (2010), cover most of the ground in terms of the statistical issues.

One side-effect of the popularity of RCT is the increasing use of Ordinary Least Squares estimators when dependant variables are binary, count or otherwise truncated in some manner. One is tempted to call this the *OLS Gone Wild* reality show, akin to the *Girls Gone Wild* reality TV show, but it is much more sober and demeaning stuff. I have long given up asking researchers in seminars why they do not just report the marginal effects for the right econometric specification. Instead I ask if we should just sack those faculty in the room who seem to waste our time teaching things like logit, count models or hurdle models. I have also volunteered that if they ever receive a referee report telling them to estimate and report the right econometric model, they can freely assume I wrote it.

Where did the notion of an RCT start? Fisher (1926) is widely acknowledged as the ‘father’ of randomisation, and indeed he did the most to systematically develop the methods. But the concept of an RCT actually originated in one of the classic debates of psychometrics: a critique by Peirce and Jastrow (1885) of the famous experiments of Fechner on subjective perceptions of differences in sensation.[9] Fechner had used his own observations of sensations to test his own theories about minimally perceptible differences, much like Fisher’s famous tea-drinking lady used cups of tea that she had prepared

[9] Regression discontinuity designs originated in psychology as well: see Thistlethwaite and Campbell (1960). Lee and Lemieux (2010) review their many applications in economics.

herself to form her opinions about the effect of having milk included before or after the tea.[10]

We often hear that an RCT is the Gold Standard in medicine, and that this should be what we unwashed social scientists should aspire to. Such claims get repeated without comment, but, to quote a popular political refrain in the United States, advocates of RCTs are entitled to their own opinions but not their own facts. Two careful studies showed that the alleged differences between an RCT and an observational study were not in fact present. Benson and Hartz (2000: 1878) '... found little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials.' Similarly, Concato, Shah and Horwitz (2000: 1887) conclude that the '... results of well-designed observational studies (with either a cohort or a case–control design) do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic'. This does not say one should not use an RCT, just that it should be used when cost-effective compared with other methods, which are often cheaper and quicker to implement.[11]

Timing is an issue that deserves more discussion. It is often difficult to design a careful RCT quickly, not because of any flaws in the method, but because of the logistical constraints of coordinating multiple sites and obtaining necessary approvals. Worrall (2007: 455–459) presents a detailed case study of a surgical procedure which was identified as being 'clearly beneficial' on the basis of observational studies, but where it took years to undertake the requisite RCT needed for the procedure to become widely recommended and used. Lives were lost because of the stubborn insistence on RCT evidence before the procedure could be widely adopted. Of course, counter-examples probably exist, but the costs and benefits of having complementary evaluation methodologies are often lost in the push to advocate one over the other.

Turning to the recent wave of applications of randomisation in economics, several concerns have been raised. Experiments are conducted to make inferences, and different types of inferences can call for different types of experiments. To take three types of inference of concern here, one might be interested in evaluating the welfare effects of a treatment for a cost–benefit analysis, one might be interested in understanding behaviour in order to design normative policies or one might be interested in estimating the

[10] Salsburg (2001) contains lively discussions of this famous anecdote, and the tensions between surrounding personalities. Hacking (1988) contains a discussion of the exotic contexts, such as the debunking of telepaths and other psychics, that led to the rise of randomisation as a popular scientific method. Of course, Fisher's humble 'seeds and soil' provided the basis for his systematic statement of the method.

[11] Prior to the popularity of RCTs, in many areas of empirical economics the typical discussion centred on the ability of weak instruments to be able to infer causality: see Rosenzweig and Wolpin (2000), Angrist and Krueger (2001), Stock, Wright and Yogo (2002) and Murray (2006). This discussion is avoided by using an RCT, although issues remain about the interpretation of causality, buried in the 'intent to treat' Sicilian defence.

(average) effects of a policy (on observables). The last of these is not usually the most important of the three.

3.2.1. Evaluating welfare effects

One can certainly be interested in worms and whatever they do, absentee teachers and whatever they do not do, the optimal use of fertiliser, wherever it comes from, savings rates and so on. But these are not substitutes for the rigorous measures of welfare from a policy, given by the equivalent variation in income. We need these measures of welfare for the application of cost–benefit analysis familiar to older generations: comparing a menu of disparate policies. How do I decide if it is better to reduce worms, increase teacher presence, use fertiliser better or increase savings rates, if I do not know the welfare impact of these policies? Of course, they might be 'costless' to implement, but that is rare.

Related to this concern, there is an important debate over the effects of charging for access to interventions. Kremer and Holla (2009) review the evidence from many RCTs in health and education that suggest that individuals and households do not seem willing to pay for interventions that generate what *seem to be* significant benefits to them at what *seem to be* significant costs. There appears to be a 'jump discontinuity' in willingness to pay that is disconcerting. At first, and second, blush this seems to be a clear revealed preference argument that the welfare benefits of the intervention are not what the researcher assumes them to be. And it leaves analysts scrambling for behavioural explanations without any empirical basis. After hand-waving about *a priori* plausible behavioural explanations, Weil (2009: 121ff) has nothing better to conclude[12] from these RCT studies than that 'the lesson here is that economists have to think more about what households know and what households think.' Is that really the best we can do?

The issue is subtle, however, as Kremer and Holla (2009) stress. Payment can change the nature of the intervention in qualitative ways, even for tiny amounts of money. An old example, from the father of field experiments, Peter Bohm, illustrates this well.[13] In 1980 he undertook a field experiment for a local government in Stockholm that was considering expanding a bus route to a major hospital and a factory. The experiment was to elicit valuations from people who were naturally affected by this route, and to test whether their aggregate contributions would make it worthwhile to provide the service. A key feature of the experiment was that the subjects would have to be willing to pay for the public good if it was to be provided for a trial period of 6 months. Everyone who was likely to contribute was given information on the experiment, but when it came time for the experiment virtually nobody turned up! The reason was that the local trade unions had decided to boycott the experiment, since it represented a threat to the current way in

[12] To be fair, he did not conduct the study and was just trying to make sense of it.

[13] Dufwenberg and Harrison (2008) provide an appreciation of the methodological significance of Bohm's pioneering work.

which such services were provided. The union leaders expressed their concerns, summarised by Bohm (1984: 136) as follows:

> They reported that they had held meetings of their own and had decided (1) that they did not accept the local government's decision not to provide them with regular bus service on regular terms; (2) that they did not accept the idea of having to pay in a way that differs from the way that 'everybody else' pays (bus service is subsidized in the area) – the implication being that they would rather go without this bus service, even if their members felt it would be worth the costs; (3) that they would not like to help in realizing an arrangement that might reduce the level of public services provided free or at low costs. It was argued that such an arrangement, if accepted here, could spread to other parts of the public sector; and (4) on these grounds, they advised their union members to abstain from participating in the project.

This fascinating outcome is actually more relevant for experimental economics in general than it might seem. When certain institutions are imposed on subjects, and certain outcomes tabulated, it does not follow that the outcomes of interest for the experimenter are the ones that are of interest to the subject. And, most critically, running field experiments forces one to be aware of the manner in which subjects select themselves into tasks based on their beliefs about the outcomes.

This process might be a direct social choice over institutions or rules, it might be Tiebout-like migration, it might be a literal or behavioural rejection of the task, it might be literal or behavioural attrition once the task is understood, it might be the evolution of social norms to resolve implicit coordination problems or it might be some combination of these. This is an active and exciting area of research in laboratory experiments now, and one that draws on insights from field experiments such as those conducted by Bohm (1984). The point is that we design better lab experiments when we worry about what one just cannot ignore in the field experiment, and those lab experiments in turn inform our inferences about the field experiment.

3.2.2. Designing normative policies

If we are to design normative policies, and understand the opportunity cost of doing so, we need to understand *why* we see certain behaviour. The apparent jump discontinuity in willingness to pay discussed above should send chills through those casually sliding from alleged 'cost effectiveness' to a recommendation that scarce resources be allocated to any project. Weil (2009) illustrates what happens when we have no complementary information on preferences or beliefs to guide our thinking. For example, consider an RCT for bed nets to prevent malaria that showed take-up rates of 40 per cent, 'even when the subsidized price is sixty cents for a bed net that lasts five years and prevents a certain number of episodes of illness or possibly death of a child' (Weil, 2009: 121). Is 40 per cent low? Who knows? Kremer and Holla (2009) and Weil (2009) think so. But here is the extent of the understanding of the issue:

- ‘Of course, any behavior can be rationalized by some combination of discount rates, value placed on child health, and so on. But it is extremely hard to do so in this case.’ (Weil, 2009: 121). How do we know it is hard to do so? Did someone ask the respondents what their time preferences were, what their subjective beliefs were, what their conditional willingness to pay for an avoided illness or even death was?
- ‘When Kremer and Holla try to think of behavioral models with some kind of procrastination going on, I become less sympathetic to their argument, partly because of the very unusual things going on here. Would the typical persons in the subject population exhibit a lot of procrastination in other aspects of life? (...) Or is this procrastination manifested only in the types of situations explored in these studies?’ (Weil, 2009: 122). Cannot we fill these massive rhetorical holes with data?
- ‘If it is the latter, that points to some other sources of the behavior, a prime candidate being some sort of information problem. That is, when I do the calculation, it is clear to me that the typical subjects in a trial should be buying this bed net for sixty cents. But maybe I have a different information structure than these persons do. Maybe they do not believe the net lasts five years, or that it works at all, or that mosquitoes cause malaria, or something like that. (. . .) Somehow these informational problems are getting tied up with the behavioral response. So I am not ready to look at the full panoply of behavioral models to rationalize this behavior.’ (Weil, 2009: 122). Huh? Subjective beliefs are not behavioural any more? And we have to wonder rhetorically about these key ingredients into the individual valuation of the mosquito-net-purchase lottery?

The frustration with this open-ended thinking comes from the knowledge that we have had the tools for a long time to answer these questions, in some measure. This type of *ex post* ‘analysis’ is like doing brain surgery with a divining rod. Or, to quote Smith (1982: 929), ‘Over twenty-five years ago, Guy Orcutt characterized the econometrician as being in the same predicament as that of an electrical engineer who has been charged with the task of deducing the laws of electricity by listening to the radio play.’

3.2.3. Evaluating intra-distributional effects

Figure 2 illustrates why we should not be lashing our inferential might to the mast of ‘the average effect’. Each panel shows the distributional impact, compared with baseline, of a policy intervention in terms of some normalised income measure. The top panel shows an average effect which is larger than the bottom panel, and would be the preferred ‘evidence based’ policy if one were to focus solely on average effects. But it has a larger standard deviation, so there are plausible levels of risk aversion that would suggest that the policy lottery with the highest average return is *not* the best one in certainty-equivalent welfare terms. Moreover, what if the welfare impact of

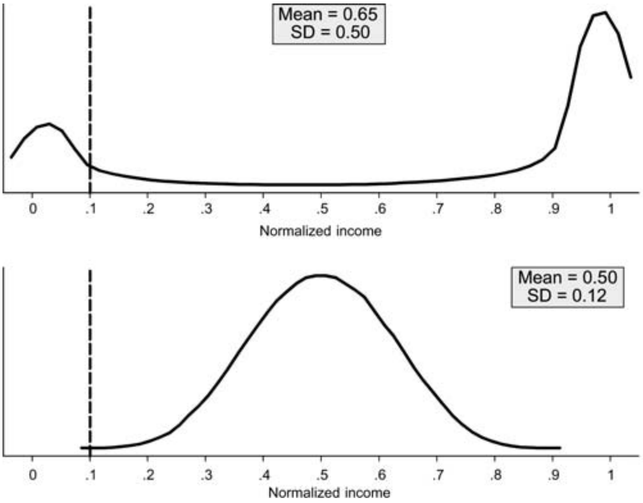

Fig. 2 Why average effects are not everything.

income levels was not uniform, such that any income level below the value of 0.1 entailed relatively high costs? Let this be an absolute poverty line, below which there is some asymmetric physiological deterioration. Then any policy that increased the chance of this outcome, even with the promise of a better income on average and even if the affected agents were risk neutral, might be a disaster. A pity, but we cannot avoid worrying about the whole distribution if we are to do a proper welfare analysis.

Of course, once one raises issues about intra-distributional effects, we can hear the Randomistas cursing those pesky unobservables, since they generate all manner of problems. Actually, they are probably just cursing Heckman (2010), or even just cursing heterogeneity itself! Anyone that does not appreciate the significance of the concern with heterogeneity should work through the arithmetic of the ‘Vietnam Draft example’ in Keane (2010: 5), and see how unreliable Wald estimators can quickly become.

The problem of randomisation bias, and the way in which it allows unobservables to affect inference, is well known. For example, when experimenters recruit subjects they offer them a lottery of earnings, offset by a fixed show-up fee. By varying the show-up fee between subjects, and measuring the risk attitudes of these that show up, one can directly demonstrate the effect of randomisation bias from this recruitment procedure (e.g. Harrison, Lau and Rutström, 2009). Turning to the RCT setting, it is well known in the field of clinical drug trials that persuading patients to participate in randomised studies is much harder than persuading them to participate in nonrandomised studies (e.g. Kramer and Shapiro, 1984). The same problem applies to social experiments, as evidenced by the difficulties that can be encountered when recruiting decentralised bureaucracies to administer the random treatment (e.g. Hotz, 1992). Heckman and Robb (1985) note that the refusal rate in one randomised job-training programme was over 90 per cent, with many of the refusals citing ethical concerns with administering a random treatment.

But apart from the statistical issues, which are bad enough, there is an important reason for wanting to keep track of the intra-distributional effects: we care a lot about 'winners' and 'losers' from policy. No policy maker can afford to ignore these equity effects, and if it is at all possible to come up with policy alternatives that mitigate losses, that is usually extremely attractive. At the very least one would like to be able to identify those individuals, and then one would like to be able to simulate policies that can mitigate losses. The simulation technology for this 'policy reform without tears' exercise is well known in trade policy evaluations, as noted earlier, but of course requires some sort of structural insight into behaviour (e.g. Harrison *et al.*, 2002a, 2004, 2003).

4. Risk and uncertainty

The evaluation of policy lotteries involves more than just the evaluation of objective risk. Even when experts are called in to offer probabilities of alternative outcomes, there is a significant element of subjectivity. Indeed, when experts are called in, without being too cynical, there is also a strategic, rent-seeking component, since experts often have a direct stake in pushing one line or the other.

One serious example, slightly stylised to protect the identity of the guilty, was the debate between 'the Americans' and 'the Europeans' over the costs of not doing anything about climate change leading up to the Kyoto negotiations. The Americans claimed that the costs of inaction were significantly smaller than the Europeans claimed. Putting aside the obvious and real political pressures for those opinions, it was easy for modellers to see where this difference came from when each side was forced to discuss the matter with numerical, structural simulation models. The European experts made extremely optimistic assumptions about a cryptic parameter known as the 'autonomous aggregate energy efficiency improvement', imaginatively denoted AAEEI. This is basically the free lunch that R&D provides in terms of the way the economy uses the available energy it has to generate output. If this is a big number, then it is easier to maintain growth with the same, or less, energy. In effect the BAU growth path gets us closer to meeting proposed Kyoto targets without tears, and without carbon taxes. But if this parameter is set to levels justified by the past decades of data, as in the simulations of the Americans, the costs of meeting the Kyoto targets are much larger, since we have to cut back growth much more than the optimistic AAEEI assumptions would suggest. Who came up with these parameter assumptions? Experts.

Does anything change when we allow for subjective beliefs in the evaluation of a policy lottery? Unfortunately, yes and no. Nothing changes if we assume, following Savage (1972), that decisions are made as if one obeys the reduction of compound lotteries (ROCL) axiom. But things change

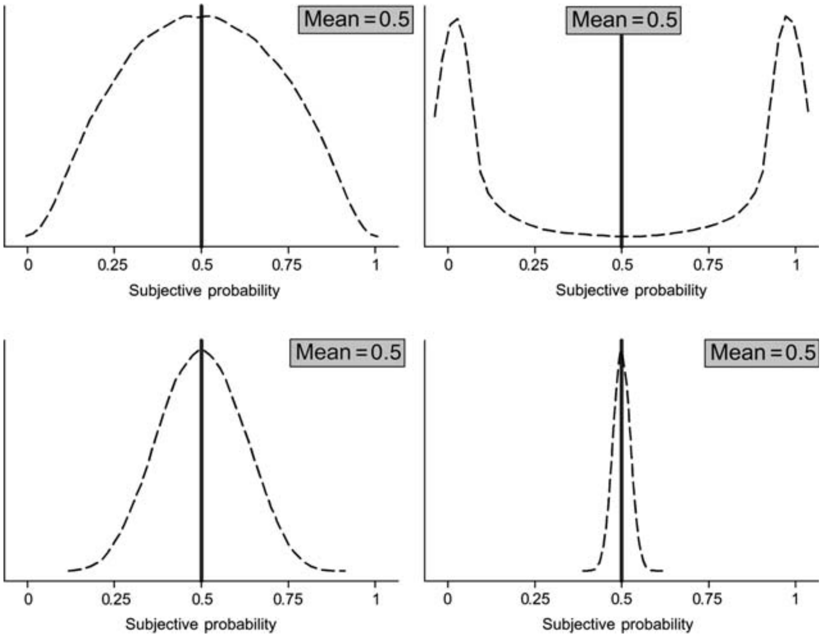

Fig. 3 Symmetric subjective probability distributions.

radically if one does not make that assumption. This seemingly technical issue is actually of great significance for the evaluation of policy lotteries, and is worth explaining carefully.

Figure 3 illustrates the situation. Assume that the subjective beliefs are symmetric, with mean one-half as shown by the solid, vertical line. But they vary in terms of the underlying distribution, as shown in the four panels of Figure 3. Some are just more or less precise than others, and one is bimodal. Under ROCL, all would generate decisions with the same outcome, since all have the same (weighted) average. Something nags at us to say that behaviour ought to be different under these different sets of beliefs, but ROCL begs to differ.

Figure 4 raises the stakes by considering asymmetric distributions. Again, ROCL is a strong, identifying assumption. Together, Figures 3 and 4 remind us that Savage (1972) did not assume that people *had* degenerate subjective probabilities that they held with certainty, he only assumed that under ROCL they behaved *as if* they did. We often forget that linguistic methodological sidestep, and confuse the 'as if' behaviour for what was actually assumed. In some cases the difference does not matter, but here it does. The reason is that when we have to worry about the underlying non-degenerate distribution, when ROCL is not assumed, then we have moved from the realm of (subjective) risk to uncertainty. And when the individual does not even have enough information to form any subjective belief distribution, degenerate or non-degenerate, we are in the realm of ambiguity.

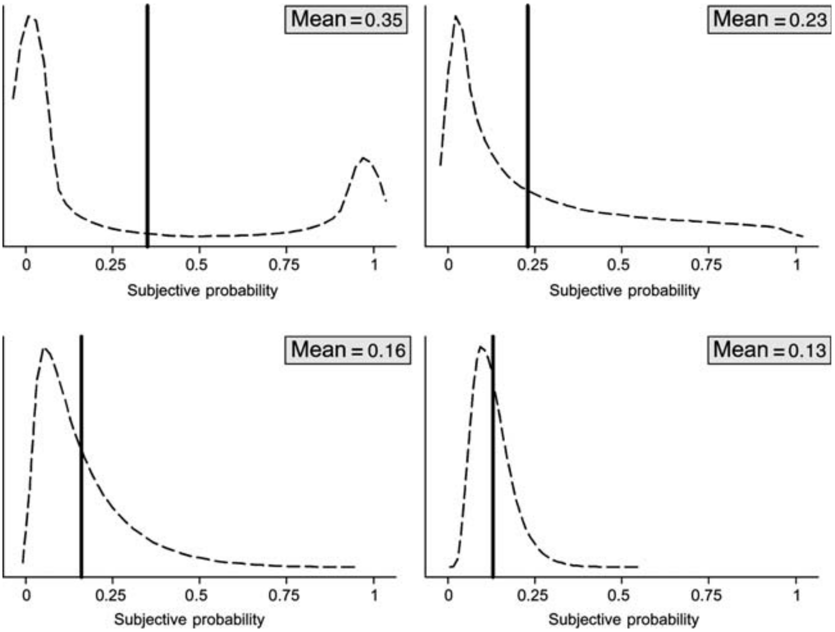

Fig. 4 Asymmetric subjective probability distributions.

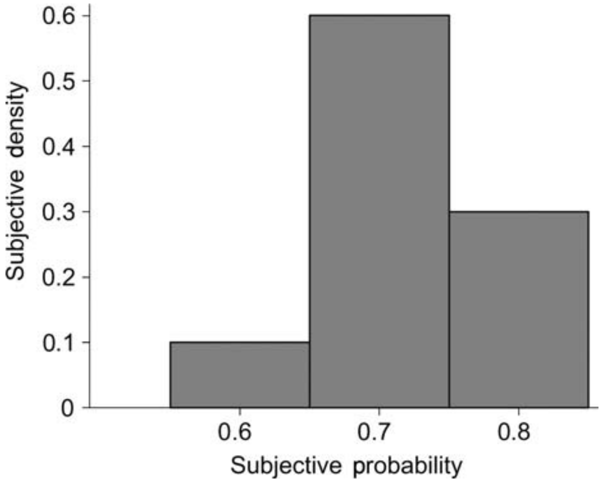

Fig. 5 ROCL at work.

Figure 5 allows a simple illustration of how ROCL allows one to collapse these disparate, non-degenerate distributions into one degenerate weighted average. Figure 5 displays a three-point discrete, non-degenerate, subjective distribution over a binary event in which the individual holds subjective probability $\pi = 0.6$ with 'prior' probability 0.1, $\pi = 0.7$ with 'prior' probability 0.6, and $\pi = 0.8$ with 'prior' probability 0.3, for a weighted average

$\pi = 0.72$. Now consider a lottery in which one gets \$X if the event occurs, and \$x otherwise. Then the subjective expected utility (SEU) is

$$0.1 \times 0.6 \times \mathrm{U(X)} + 0.1 \times 0.4 \times \mathrm{U(x)} + 0.6 \times 0.7 \times \mathrm{U(X)} + 0.6 \times 0.3 \times \mathrm{U(x)} + 0.3 \times 0.8 \times \mathrm{U(X)} + 0.3 \times 0.2 \times \mathrm{U(x)},$$

which collapses to

$$(0.1 \times 0.6 + 0.6 \times 0.7 + 0.3 \times 0.8) \times \mathrm{U(X)} + (0.1 \times 0.4 + 0.6 \times 0.3 + 0.3 \times 0.2) \times \mathrm{U(x)}$$

and hence to

$$0.72 \times \mathrm{U(X)} + 0.28 \times \mathrm{U(x)}$$

under ROCL. So the non-degenerate distribution in Figure 5 can be boiled down to a degenerate subjective probability of 0.72 under ROCL: an impressive identifying restriction!

How we relax ROCL is a matter for important, foundational research. Although it has taken half a century for the implications of Ellsberg (1961) to be formalised in tractable ways, we are much closer to doing so. One popular approach is the 'smooth ambiguity model' of Klibanoff, Marinacci and Mukerji (2005), with important parallels in Davis and Paté-Cornell (1994), Ergin and Gul (2009), Nau (2006) and Neilsen (2010). Another popular approach is due to Ghirardoto, Maccheroni and Marinacci (2004), generalising Gilboa and Schmeidler (1989).

We can illustrate the smooth ambiguity model with a simple example. Let $\mathrm{CE}(\pi = 0.6)$ be the certainty equivalent of the lottery $0.6 \times \mathrm{U(X)} + 0.4 \times \mathrm{U(x)}$, $\mathrm{CE}(\pi = 0.7)$ be the certainty equivalent of the lottery $0.7 \times \mathrm{U(X)} + 0.3 \times \mathrm{U(x)}$, and $\mathrm{CE}(\pi = 0.8)$ be the certainty equivalent of the lottery $0.8 \times \mathrm{U(X)} + 0.2 \times \mathrm{U(x)}$. Then the evaluation of the lottery can be written

$$0.1 \times \phi(\mathrm{CE}(\pi = 0.6)) + 0.6 \times \phi(\mathrm{CE}(\pi = 0.7)) + 0.3 \times \phi(\mathrm{CE}(\pi = 0.8)),$$

where ϕ is a function defined over the certainty-equivalent of the lottery that is conditional on a particular subjective probability value. Akin to the properties of $\mathrm{U}(\cdot)$ defining risk attitudes under EUT or SEU, the properties of $\phi(\cdot)$ define attitudes towards the uncertainty over the particular subjective probability value.[14] If ϕ is concave, then the decision-maker is uncertainty averse; if ϕ is convex, then the decision-maker is uncertainty loving and if ϕ is linear,

[14] In the original specifications ϕ is said to characterise attitudes towards ambiguity, but the earlier definition of risk, uncertainty and ambiguity makes it apparent why one would not want to casually confound the two. One would only be dealing with ambiguity in the absence of well-defined prior probabilities over the three subjective probability values 0.6, 0.7 and 0.8.

then the decision-maker is uncertainty neutral. The familiar SEU specification emerges if ϕ is linear, since then ROCL applies after some irrelevant normalisation. The overall evaluation of the lottery depends on risk attitudes *and* uncertainty attitudes, and there is no reason for the decision-maker to be averse to both at the same time. An important econometric corollary is that one cannot infer attitudes towards uncertainty from observed choice until attitudes towards risk are characterised.

5. Implications

We now have many rich models of behaviour, allowing structural understanding of decisions in many settings of interest for the design of agricultural, food and resource policy. But we also realise that there are some basic confounds to reliable inference about behaviour. These are not side technical issues. Risk attitudes can involve more than diminishing marginal utility, and we have no significant problems identifying alternative paths to risk aversion through probability weighting. Loss aversion is much more fragile, until we can claim to know the appropriate reference points for agents. Time preferences can be characterised, and appear to hold fewer problems than early experimental studies with lab subjects suggest.

But the 600 pound gorilla confound is the subjective belief that decision-makers hold in many settings. This is the one that is widely ignored. The suggestion is not that it should be used to rationalise 'rational behaviour' in every setting, but that inferences about cognitive failures, and the need for nudges, hinge on our descriptive knowledge of what explains behaviour. If we rule out some factor, then something else may look odd.[15] Of course, in some settings it is simply not possible to 'go back to the well' and elicit information of this kind. But there is no reason why one cannot use information from one sample, even from a different population if necessary, to condition inferences about another sample, to see the effect.[16]

These preferences and beliefs have been elicited reliably in lab settings and in the field, although the myriad of contexts of the field mean that each application is in some important sense unique. The question to be asked is why these methods are not used more frequently in RCT evaluations of policies. This is beginning, but the attempts to elicit preferences and beliefs in existing randomised evaluations have been casual at best. Here we have a hypothetical survey question about risky behaviour, there we have an unmotivated question

[15] To take a simple example, assume that there is a risk premium, but one uses either a model that assumes that 100 per cent of the observed behaviour is due to diminishing marginal utility or a model that assumes that 100 per cent of the observed behaviour is due to probability pessimism. The first model will generate concave utility functions, and the second model will generate convex probability weighting functions: both will likely explain behaviour tolerably well.

[16] Coller, Harrison and Rutström (2011) provide an example in which estimates of the utility function were generated from choices made by one sample from a population, and then used to condition inferences about discount rates from another sample from the same population. Although second-best, there is no econometric reason one cannot undertake inferences in this manner when the first-best option is unavailable or too costly.

about beliefs and rarely do we try to elicit time preferences at all. The potential complementarity between these methods is obvious, and conceded by all, but there seems to be relatively little appetite for careful field experiments to elicit preferences and beliefs. In part this derives from the way in which randomised evaluations have been marketed and promoted intellectually, as an antidote to the need to make structural economic or econometric assumptions.

The next generation of field experiments will illustrate the value of combining tasks that allow one to estimate latent structural parameters with interventions that allow the sharp contrast between control and treatment. The next generation of econometric analysts will use the insights from these structural models to inform their understanding of the distributional impacts of interventions, rather than just the average impact.[17] They will also use these structural parameters to gauge the sample selection issues that plague randomised interventions of sentient objects, rather than agricultural seeds. And both groups of researchers will find themselves heading back to the lab to validate their experimental designs and econometric methods applied to field data. There they will find time to talk to theorists again, who have produced some beautiful structures needed to help understand subjective risk and uncertainty.

Acknowledgements

I am grateful to Harold Kincaid for discussions of the literature.

References

Andersen, S., Fountain, J., Harrison, G. W. and Rutström, E. E. (2010). Estimating subjective probabilities. Working Paper 2010–06, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Andersen, S., Fountain, J., Harrison, G. W., Hole, A. R. and Rutström, E. E. (2011c). Inferring beliefs as subjectively imprecise probabilities. *Theory and Decision*, forthcoming.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. E. (2008). Eliciting risk and time references. *Econometrica* 76(3): 583–619.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. E. (2011a). Discounting behavior: a reconsideration. Working Paper 2011–03, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. E. (2011b). Intertemporal utility and correlation aversion. Working Paper 2011–04, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

[17] This is not the same thing as saying that they will build full structural models of the effect of the intervention, although this is not ruled out. Advocates of randomised interventions often pose a false dichotomy between 'all-in theological' 'modeling via structural assumptions' or 'agnostic eyeballing' of the average effects: Heckman (2010), in particular, takes aim squarely at this false tradeoff. The former is very hard to do well, and quite easy to do poorly. The latter is fine as far as it goes, but just does not go very far.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives* 15(4): 69–85.

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Banerjee, A. V. and Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics* 1: 151–178.

Benson, K. and Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine* 342(25): 1878–1886.

Bohm, P. (1984). Are there practicable demand-revealing mechanisms? In: H. Hanusch (ed.), *Public Finance and the Quest for Efficiency*. Detroit: Wayne State University Press.

Coller, M., Harrison, G. W. and Rutström, E. E. (2011). Latent process heterogeneity in discounting behavior. *Oxford Economic Papers*, forthcoming.

Coller, M. and Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics* 2: 107–127.

Concato, J., Shah, N. and Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine* 342(25): 1887–1892.

Davis, D. B. and Paté-Cornell, M. E. (1994). A challenge to the compound lottery axiom: A two-stage normative structure and comparison to other theories. *Theory and Decision* 37: 267–309.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature* 48(2): 424–455.

Desvousges, W. H., Johnson, F. R. and Banzhaf, H. S. (1999). *Environmental Policy Analysis with Limited Information: Principles and Applications of the Transfer Method*. New York: Elgar.

Duflo, E. (2006). Field experiments in development economics. In: R. Blundell, W. Newey and T. Persson (eds), *Advances in Economics and Econometrics: Theory and Applications*, Vol. 2. New York: Cambridge University Press.

Duflo, E., Glennerster, R. and Kremer, M. (2007). Using randomization in development economics research: A toolkit. In: T. P. Schultz and J. Strauss (eds), *Handbook of Development Economics*, Vol. 2. New York: North-Holland.

Duflo, E. and Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. In: G. Pitman, O. Feinstein and G. Ingram (eds), *Evaluating Development Effectiveness*. New Brunswick, NJ: Transaction Publishers.

Dufwenberg, M. and Harrison, G. W. (2008). Peter Bohm: father of field experiments. *Experimental Economics* 11(3): 213–220.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Ergin, H. and Gul, F. (2009). A theory of subjective compound lotteries. *Journal of Economic Theory* 144(3): 899–929.

Fiore, S. M., Harrison, G. W., Hughes, C. E. and Rutström, E. E. (2009). Virtual experiments and environmental policy. *Journal of Environmental Economics & Management* 57(1): 65–86.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture* 33(1): 503–513.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time discounting and time preference: a critical review. *Journal of Economic Literature* 40: 351–401.

Ghirardoto, P., Maccheroni, F. and Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* 118: 133–173.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18: 141–153.

Hacking, I. (1988). Telepathy: origins of randomization in experimental design. *Isis* 79: 427–451.

Harrison, G. W. (2006). Experimental evidence on alternative environmental valuation methods. *Environmental and Resource Economics* 34: 125–162.

Harrison, G. W., Humphrey, S. J. and Verschoor, A. (2010). Choice under uncertainty: evidence from Ethiopia, India and Uganda. *Economic Journal* 120: 80–104.

Harrison, G. W., Jensen, S. E., Pedersen, L. and Rutherford, T. F. (eds) (2000). *Using Dynamic General Equilibrium Models for Policy Analysis*. Amsterdam: Elsevier, Contributions to Economics Analysis 248.

Harrison, G. W., Jensen, J., Lau, M. I. and Rutherford, T. F. (2002a). Policy reform without tears. In: A. Fossati and W. Weigard (eds), *Policy Evaluation with Computable General Equilibrium Models*. New York: Routledge.

Harrison, G. W., Lau, M. I. and Rutström, E. E. (2007). Estimating risk attitudes in Denmark: a field experiment. *Scandinavian Journal of Economics* 109(2): 341–368.

Harrison, G. W., Lau, M. I. and Rutström, E. E. (2009). Risk attitudes, randomization to treatment, and self-selection into experiments. *Journal of Economic Behavior and Organization,* 70(3): 498–507.

Harrison, G. W., Lau, M. I. and Rutström, E. E. (2011). Theory, experimental design and econometrics are complementary (and so are lab and field experiments). In: G. Frechette and A. Schotter (eds), *The Methods of Modern Experimental Economics*. New York: Oxford University Press, forthcoming.

Harrison, G. W., Lau, M. I. and Williams, M. B. (2002b). Estimating individual discount rates for Denmark: a field experiment. *American Economic Review* 92(5): 1606–1617.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature* 42(4): 1013–1059.

Harrison, G. W., List, J. A. and Towe, C. (2007). Naturally occurring preferences and exogenous laboratory experiments: a case study of risk aversion. *Econometrica* 75(2): 433–458.

Harrison, G. W., Rutherford, T. F. and Tarr, D. G. (2003). Trade liberalization, poverty and efficient equity. *Journal of Development Economics* 71: 97–128.

Harrison, G. W., Rutherford, T. F., Tarr, D. G. and Gurgel, A. (2004). Trade policy and poverty reduction in Brazil. *World Bank Economic Review* 18(3): 289–317.

Harrison, G. W. and Rutström, E. E. (2008). Risk aversion in the laboratory. In: J. C. Cox and G. W. Harrison (eds), *Risk Aversion in Experiments*, Vol. 12. Bingley, UK: Emerald, Research in Experimental Economics.

Harrison, G. W. and Rutström, E. E. (2009). Expected utility *and* prospect theory: one wedding and a decent funeral. *Experimental Economics,* 12(2): 133–158.

Harrison, G. W. and Vinod, H. D. (1992). The sensitivity analysis of applied general equilibrium models: completely randomized factorial sampling designs. *Review of Economics and Statistics* 74: 357–362.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature* 48(2): 356–398.

Heckman, J. J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In: J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press.

Hotz, V. J. (1992). Designing an evaluation of JTPA. In: C. Manski and I. Garfinkel (eds), *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.

Imbens, G. W. (2010). Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48(2): 399–423.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics* 156: 3–20.

Klibanoff, P., Marinacci, M. and Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica* 73(6): 1849–1892.

Kremer, M. and Holla, A. (2009). Pricing and access: lessons from randomized evaluations in education and health. In: W. Easterly and J. Cohen (eds), *What Works in Development: Thinking Big and Thinking Small*. Washington DC: Brookings Institution Press.

Kramer, M. and Shapiro, S. (1984). Scientific challenges in the application of randomized trials. *Journal of the American Medical Association,* 252(19): 2739–2745.

Lau, M. I. (2000). Assessing tax reforms when human capital is endogenous. In: G. W. Harrison, S. E. H. Jensen, L. H. Pedersen and T. F. Rutherford (eds), *Using Dynamic General Equilibrium Models for Policy Analysis*. Amsterdam: North Holland, Contributions to Economic Analysis 248.

Leamer, E. E. (2010). Tantalus on the road to Asymptopia. *Journal of Economic Perspectives,* 24(2): 31–46.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2): 281–355.

Mathieson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* 22(10): 1087–1096.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* 20(4): 111–132.

Nau, R. F. (2006). Uncertainty aversion with second-order utilities and probabilities. *Management Science* 52: 136–156.

Neilson, W. S. (2010). A simplified axiomatic approach to ambiguity aversion. *Journal of Risk and Uncertainty* 41: 113–124.

Peirce, C. S. and Jastrow, J. (1885). On small differences of sensation. *Memoirs of the National Academy of Sciences for 1884,* 3: 75–83.

Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural 'natural experiments' in economics. *Journal of Economic Literature* 38: 827–874.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Freeman.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of American Statistical Association* 66: 783–801.

Savage, L. J. (1972). *The Foundations of Statistics*, 2nd edn. New York: Dover Publications.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *American Economic Review,* 72(5): 923–955.

Sorenson, R. A. (1992). *Thought Experiments*. New York: Oxford University Press.

Stern, N. (2007). *The Economics of Climate Change: The Stern Review*. New York: Cambridge University Press.

Stock, J. H., Wright, J. H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20(4): 518–529.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: an alternative to the ex post facto experiment. *Journal of Educational Psychology* 51(6): 309–317.

Weil, D. N. (2009). Comment. In: W. Easterly and J. Cohen (eds), *What Works in Development: Thinking Big and Thinking Small*. Washington DC: Brookings Institution Press.

Worrall, J. (2007). Why there's no cause to randomize. *British Journal of the Philosophy of Science* 58: 451–488.