8-1-2010

Theory, Experimental Design and Econometrics Are Complementary (And So Are Lab and Field Experiments)

Glenn Harrison
Georgia State University

Morten Lau
University of Newcastle

Elisabet Rutström
Georgia State University

Theory, Experimental Design and Econometrics

Are Complementary (And So Are Lab and Field Experiments)

by

Glenn W. Harrison, Morten Lau & E. Elisabet Rutström[†]

August 2010

ABSTRACT. Experiments are conducted with various purposes in mind including theory testing, mechanism design and measurement of individual characteristics. In each case a careful researcher is constrained in the experimental design by prior considerations imposed either by theory, common sense or past results. We argue that the integration of the design with these elements needs to be taken even further. We view all these elements that make up the body of research methodology in experimental economics as mutually dependant and therefore take a systematic approach to the design of our experimental research program. Rather than drawing inferences from individual experiments or theories as if they were independent constructs, and then using the findings from one to attack the other, we recognize the need to constrain the inferences from one by the inferences from the other. Any data generated by an experiment needs to be interpreted jointly with considerations from theory, common sense, complementary data, econometric methods and expected applications. We illustrate this systematic approach by reference to a research program centered on large artefactual field experiments we have conducted in Denmark. An important contribution that grew out of our work is the complementarity between lab and field experiments.

[†] Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison); Newcastle University Business School, University of Newcastle (Lau); and Robinson College of Business, Georgia State University, USA (Rutström). E-mail contacts: gharrison@gsu.edu, morten.lau@newcastle.ac.uk and erutstrom@gmail.com. Steffen Andersen and Melonie Sullivan have made significant contributions to the research discussed here. We thank the U.S. National Science Foundation for research support under grants NSF/HSD 0527675 and NSF/SES 0616746, and the Danish Social Science Research Council for research support under projects 24-02-0124 and 275-08-0289.

Table of Contents

1. Policy Lotteries . . . . . -4-

2. Risk Aversion . . . . . -10-
A. Sampling Procedures . . . . . -12-
B. Elicitation Procedures . . . . . -14-
C. Estimation Procedures . . . . . -18-
D. Asset Integration . . . . . -26-

3. Discount Rates . . . . . -29-
A. Defining Discount Rates in Terms of Utility . . . . . -30-
B. The Need for Joint Estimation . . . . . -32-

4. Lessons Learned . . . . . -33-
A. The Role of Artefactual Field Experiments . . . . . -34-
B. The Contrived Debate Between Lab and Field Experiments . . . . . -36-
C. Danes Are Like Plain Yogurt, Not Like Wines Or Cheeses . . . . . -37-
D. Non-EUT Models of Risky Choice and Non-Exponential Models of Discounting: *Festine Lente* . . . . . -38-
E. Estimation, Not Direct Elicitation . . . . . -41-
F. Virtual Experiments As A Smooth Bridge Between the Lab and the Field . . . . . -42-

5. Conclusions . . . . . -46-

References . . . . . -52-

Experiments are conducted with various purposes in mind including theory testing, mechanism design and measurement of individual characteristics. In each case a careful researcher is constrained in the experimental design by prior considerations imposed either by theory, common sense or past results. We argue that the integration of the design with these elements needs to be taken even further. We view all these elements that make up the body of research methodology in experimental economics as mutually dependant and therefore take a systematic approach to the design of our experimental research program. Rather than drawing inferences from individual experiments or theories as if they were independent constructs, and then using the findings from one to attack the other, we recognize the need to constrain the inferences from one by the inferences from the other. Any data generated by an experiment needs to be interpreted jointly with considerations from theory, common sense, complementary data, econometric methods and expected applications.

We illustrate this systematic approach by reference to a research program centered on large artefactual field experiments we have conducted in Denmark.[1] The motivation for our research was

[1] There is a growing literature of experiments performed outside of university research laboratories, building on the pioneering work of Peter Bohm over many years, starting in the 1970s. Dufwenberg and Harrison [2008; p.214ff.] provide a posthumous appreciation of his motivation: "Peter was drawn to conduct field experiments long before laboratory experiments had become a staple in the methodological arsenal of economists. Just as some experimentalists do not comprehend why one would ask questions with no real economic consequences, or care too much about the responses to such questions, Peter began doing field experiments simply because they answered the questions he was interested in. He did not come to field experiments because of any frustration with lab experiments, or from any long methodological angst about laboratory experiments: it was just obvious to him that experiments needed field referents to be interesting. He later became interested in the methodological differences between laboratory and field experiments, well after his own pioneering contributions to the later had been published." Due to the great variety of such experiments with respect to procedures, contexts and participant pools there has been a refinement of the field-lab terminology to include modifiers such as "artefactual." We will restrict our discussions to two kinds of experiments only: the traditional research laboratory using convenient and low cost student samples, and the artefactual field experiment that employs samples from populations not restricted to students. In these latter experiments the tasks are similar to those presented to students but often have to be adjusted to the perceptual and conceptual needs of the subject pool. Here we will simply use the label "lab" when referring to experiments we conduct on student samples and "field" to those conducted on samples from more heterogeneous field populations.

to generate measures of household and individual characteristics for use in a range of policy valuations. An important contribution that grew out of our work is the complementarity of lab and field experiments.

One such characteristic was the risk preferences of representative Danish residents. Predicted welfare effects from policy changes are always uncertain, in part because of imprecisely known parameter values in the policy simulation models used. We introduce the term "policy lottery" to refer to such uncertainties over the predicted policy effects. In light of these uncertainties, we argue that the welfare impact calculated for various households should reflect their risk attitudes. When comparing policies with similar expected benefits but with differences in the uncertainty over those predicted effects, a risk averse household would prefer the policy with less uncertain effects to that with more. Including measures of risk attitudes in policy evaluations can therefore have important implications for inferences about the distribution of welfare effects. This is a significant improvement over the standard practice in policy evaluations that either assume risk neutrality or some arbitrarily selected risk coefficient employed uniformly over all household types. Our dominating justification for the expense of going out in the field derived from the policy need to provide measures for households and individuals that are representative of the general Danish population.

An instrumental part of our research program was the inclusion of a number of complementary lab experiments conducted at a much lower cost because of the use of convenience subject pools: students. Due to the lower cost we could conduct a wider range of robustness tests varying elicitation instruments and procedures, but because we sampled from a more restricted population these results are not by themselves informative to the policy applications we have in mind. Nevertheless, the results obtained from such convenience samples can be used to condition the inferences drawn from the observations on the field sample.

Another aspect of the systematic approach was to use several theoretical considerations to guide our experimental design from the start. One important characteristic that we measure is the discount rate of individuals across various household types. Theory is quite clear that what is being discounted is not the money stream but the stream of utility that derives from that money. Recognition of this fact had an influence not only on the inclusion of tasks incorporating both risk and time manipulations but also on the econometric strategy of joint estimation. The joint estimation approach leads to estimates of risk attitudes that are consistent with the estimated discount rates and *vice versa.*

Finally, the systematic approach we advocate encourages the use of common sense constraints on the inferences drawn from the data. For example, many structural model specifications suffer from inflexibility globally so that they provide poor predictions on domains outside the one on which the data was generated. The Constant Relative Risk Aversion function, for example, if estimated on small stakes can make predictions on large stakes that may appear ridiculous. The same may even be the case for the more flexible Expo-power function if estimated on a stake domain where the income effect is negligible. Inferences drawn from estimations using restrictive domains and restrictive specifications must therefore be constrained with common sense constraints on their applicability.

In section 1 we introduce the concept of policy lotteries, giving a few examples. In section 2 we discuss how we draw inferences about risk attitudes using our systematic approach that includes conditioning these inferences on smaller scale lab experiments, on sample selection effects and elicitation methods, on econometric and statistical strategies such as sampling frame and structural estimation approaches, and on theoretical and common sense considerations about out-of-domain predictions. In section 3 we discuss inferences about discount rates and demonstrate the power of joint estimation of risk and time preferences as motivated by theory. Section 4 expands the joint

inference discussion to longitudinal issues such as temporal stability.

1. Policy Lotteries

The motivation for the field experiments on which this research program is centered came from our earlier work with the Danish Ministry of Business and Industry between 1996 and 2000 to develop computable general equilibrium (CGE) models of public policy. Those policies ranged from general tax reforms to specific carbon tax reforms, from the effects of relaxing domestic retail opening hours to the effects on Denmark of global trade reform, from intergenerational welfare issues to the dynamics of human capital formation. One of the hallmarks of the CGE models we were developing 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, Jensen, Pedersen and Rutherford [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. We elicited discount rates, in part, to address that policy question directly (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. Note that this is uncertainty in the *estimate of the impact*, and not necessarily uncertainty in the *impact itself*. But we submit that in the limited information world of practical policy-making such uncertainties are rife.[2]

We illustrate the concept of a policy lottery using the CGE model documented in Harrison, Jensen, Lau and Rutherford [2002]. 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.[3] 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. More elaborate versions of the model exist in which inter-temporal and inter-generational behavior are modeled (e.g., Lau [2000]), but this static version is ideal for our illustrative purposes.

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 7 households: singles younger than 45 without children, singles older than 45 without children, households younger than 45 without

[2] For example, see Desvousges et al. [1999]. The limitation on information can derive from the inherent difficulty of modeling behavioral or physical relationships, from the short time-frame over which the model has to be developed and applied, or both.

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

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 randomized 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.[4] 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

[4] 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% 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.

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 interquartile range,[5] 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.

Generation of policy lotteries are not restricted to CGE models. The method applies to any simulation model that generates outcomes that reflect policy changes. For example, Fiore, Harrison, Hughes and Rutström [2009] used a simulation model of the spread of forest fire, developed by the USDA for that purpose and calibrated to detailed GIS data for a specific area, to generate policy lotteries for experimental subjects to make choices over. Our approach just recognizes that policy models of this kind are never certain, and that they contain standard errors: in fact lots of standard errors. But that uncertainty should not be ignored when the policy maker uses the model to decide on good policies.

The idea that policies are lotteries is a simple one, and well known in the older simulation literature in CGE modeling. The methods developed to address it amounted to Monte Carlo analyses on repeated simulations in which each uncertain parameter was perturbed around its point estimate. By constraining these perturbations to within some empirical or *a priori* confidence region, one implicitly constrained the simulated policy outcome to that region. The same idea plays a central role in the *Stern Review on the Economics of Climate Change* (Stern [2007]). It stresses (p.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

[5] Defined by the 25th and 75th percentiles, this range represents 50% of the observations around the median.

utility analysis" of the costs of climate change (p. 173ff.) which is effectively the same as viewing climate change impacts as 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.

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 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 our experimental results that these households are not risk neutral. Assume a Constant Relative Risk Aversion (CRRA) utility specification for each household. Anticipating the later discussion of our experimental results, we can stratify our raw elicited CRRA intervals according to these 7 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%, 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%, and that the government affected lump-sum side payments to each household to ensure that no household had less than a 1% welfare gain.[6] In this case, plausible elasticity configurations for the model exist that result in very large welfare gains for some households.[7] 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% 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. Given the evidence from our estimation of the Expo-Power function, reported in Harrison, Lau and Rutström [2006], 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 7 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 computable general equilibrium 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, Rutherford and

[6] The manner in which these sidepayments are computed is explained in Harrison, Jensen, Lau and Rutherford [2002]. It corresponds to a stylized version of the type of political balancing act one often encounters behind the scenes in the design of a public policy such as this.

[7] 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.

Tarr [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 generalize from convenience samples. In a related setting, in this instance with respect to behavioral findings from laboratory experiments that question some of the theoretical foundations of welfare economics, List [2005; p.36] records that in his

> ... discussions with agency officials in the U.S. who perform/oversee benefit-cost analyses, many are aware of these empirical findings, and realize that they have been robust across unfamiliar goods, such as irradiated sandwiches, and common goods, such as chocolate bars, but many remain skeptical of the received results. Most importantly for our purposes, some policymakers view experimental laboratory results with a degree of suspicion, one noting that the methods are akin to "scientific numerology." When pressed on this issue, some suggest that their previous experience with stated preference surveys leads them to discount experimental results, especially those with student samples, and they conclude that the empirical findings do not merit policy changes yet. A few policy officials openly wondered if the anomalous findings would occur in experiments with "real" people.

Our experience has been the same, and is why we were led to conduct field experiments in Denmark.

2. Risk Aversion

In order to evaluate the policy lottery considered in the previous section, we needed to have estimates of the risk attitudes for the different households in Denmark. We therefore designed an experiment to elicit risk attitudes (and discount rates) from representative Danes. The experiment is a longitudinal panel where we revisited many of the first stage participants at a later date. In this

section we discuss the issues that arose in our field experiments, with an emphasis on those issues that are relatively novel as a result of the field context.

The immediate implication, of course, was that we needed to generate a sampling frame that allowed us to make inferences about the broader adult population in Denmark. This led us to employ stratified sampling methods for large-scale surveys, which are relatively familiar to labor economists and health economists, but which had not been used in the experimental literature. We were also concerned about possible sample selection effects from our recruiting strategy, and the possibility of what is known in the literature as "randomization bias." Two types of sample selection effects were possible. First, we were concerned that the information about earnings in the recruitment information would attract a sample biased in the direction of risk loving. Second, we were concerned that particular experiences in the first stage of the experiment could bias attrition to the second stage. These concerns influenced not only our econometric strategy but also lead to the design of complementary lab experiments to directly test for such effects.

The next concern was with the design of the elicitation procedure itself. There were many alternatives available in the literature, and known trade-offs from using one or the other. We were particularly concerned to have an elicitation procedure that could be relatively easily implemented in the field, even though we had the benefit compared to some field contexts of being able to assume a literate population. We use elicitation procedures that do not have a specific context since the purpose was to generate risk preference parameters for general policy use. We use complementary lab experiments to condition our field inferences on any vulnerability in responses to variations in procedures. These procedural variations were guided by hypotheses about the effect of frames on the participants' perception of the task and on their use of information processing heuristics.

Once we had collected the experimental data, several issues arose concerning the manner in which one infer the risk attitudes. These issues demanded the use of an explicit, structural approach

to estimating models of choice over risky lotteries. The reason is that we wanted to obtain estimates of the latent parameters of these choice models, and to be able to evaluate alternative choice models at a structural level. One attraction of this approach is that it allowed us to be explicit about issues that are often left implicit, but which can have a dramatic affect on inferred risk attitudes; one example is the specification of what is known as a "behavioral error term" in these choice models. Another attraction is that it allowed us to examine alternative theories to expected utility theory using a comparable inferential framework.

Our goal was to generate measures of risk attitudes for a range of monetary prizes and over time. With this data we can investigate the robustness of the measures over time, as reflective of stationary or state dependent preferences, and robustness with respect to income changes.

A. Sampling Procedures

The sample for the field experiments was designed to be representative of the adult Danish population in 2003. There were six steps in the construction of the sample, detailed in Harrison, Lau, Rutström and Sullivan [2005] and essentially following those employed in Harrison, Lau and Williams [2002]:

- First, a random sample of 25,000 Danes was drawn from the Danish Civil Registration Office in January 2003. Only Danes born between 1927 and 1983 were included, thereby restricting the age range of the target population to between 19 and 75. For each person in this random sample we had access to their name, address, county, municipality, birth date, and sex. Due to the absence of names and/or addresses, 28 of these records were discarded.
- Second, we discarded 17 municipalities (including one county) from the population, due to them being located in extraordinarily remote locations, and hence being very costly to recruit. The population represented in these locations amounts to less than 2% of the

Danish population, or 493 individuals in our sample of 25,000 from the Civil Registry. Hence it is unlikely that this exclusion could quantitatively influence our results on sample selection bias.

- Third, we assigned each county either 1 session or 2 sessions, in rough proportionality to the population of the county. In total we assigned 20 sessions. Each session consisted of two sub-sessions at the same locale and date, one at 5pm and another at 8pm, and subjects were allowed to choose which sub-session suited them best.
- Fourth, we divided 6 counties into two sub-groups because the distance between some municipalities in the county and the location of the session would be too large. A weighted random draw was made between the two sub-groups and the location selected, where the weights reflect the relative size of the population in September 2002.
- Fifth, we picked the first 30 or 60 randomly sorted records within each county, depending on the number of sessions allocated to that county. This provided a sub-sample of 600.
- Sixth, we mailed invitations to attend a session to the sub-sample of 600, offering each person a choice of times for the session. Response rates were low in some counties, so another 64 invitations were mailed out in these counties to newly drawn subjects.[8] Everyone that gave a positive response was assigned to a session, and our recruited sample was 268.

Attendance at the experimental sessions was extraordinarily high, including 4 persons who did not respond to the letter of invitation but showed up unexpectedly and participated in the experiment. Four persons turned up for their session, but were not able to participate in the

[8] We control for county and the recruitment wave to which the subject responded in our statistical analysis of sample selection. Response rates were higher in the greater Copenhagen area compared to the rest of the country. The experiments were conducted under the auspices of the Ministry of Economic and Business Affairs, and people living outside of the greater Copenhagen area may be suspicious of government employees and therefore less likely to respond to our letter of invitation.

experiments.[9] These experiments were conducted in June of 2003, and a total of 253 subjects participated.[10] Sample weights for the subjects in the experiment can be constructed using this experimental design, and can be used to calculate weighted distributions and averages that better reflect the adult population of Denmark.

B. Elicitation Procedures

There are many general elicitation procedures that have been used in the literature to ascertain risk attitudes from individuals in the experimental laboratory using non-interactive settings, and each is reviewed in detail by Harrison and Rutström [2008]. Most of these simply present participants with lotteries specified using various monetary prizes and probabilities without attaching a particular context: these are labeled "artefactual" presentations. An approach made popular by Holt and Laury [2002] is the Multiple Price List (MPL), which entails giving the subject an ordered array of binary lottery choices to make all at once. The MPL requires the subject to pick one of the lotteries on offer, and then plays that lottery out for the subject to be rewarded. The earliest use of the MPL design in the context of elicitation of risk attitudes is, we believe, Miller, Meyer and Lanzetta [1969]. Their design confronted each subject with 5 alternatives that constitute an MPL, although the alternatives were presented individually over 100 trials. The method was later used by Schubert, Brown, Gysler and Brachinger [1999], Barr and Packard [2002] and, of course, Holt and

[9] The first person suffered from dementia and could not remember the instructions; the second person was a 76 year old woman who was not able to control the mouse and eventually gave up; the third person had just won a world championship in sailing and was too busy with media interviews to stay for two hours; and the fourth person was sent home because they arrived after the instructions had begun and we had already included one unexpected "walk-in" to fill their position.

[10] Certain events might have plausibly triggered some of the no-shows: for example, 3 men did not turn up on June 11, 2003, but that was the night that the Danish national soccer team played a qualifying game for the European championships against Luxembourg that was not scheduled when we picked session dates.

Laury [2002]. The MPL has the advantage of allowing the subject to easily compare options involving various risks. As is the case with all procedures of this nature there is some question about the robustness of responses with respect to procedural variations. We decided to use complementary lab experiments to explore several of these procedural issues, rather than incur the expense of evaluating them in the field.

In our field version of the MPL each subject is presented with a choice between two lotteries, which we can call A or B. Table 1 illustrates the basic payoff matrix presented to subjects in our experiments. The complete procedures are described in Harrison, Lau, Rutström and Sullivan [2005]. The first row shows that lottery A offered a 10% chance of receiving 2,000 DKK and a 90% chance of receiving 1,600 DKK. The expected value of this lottery, EV^A, is shown in the third-last column as 1,640 DKK, although the EV columns were not presented to subjects. Similarly, lottery B in the first row has chances of payoffs of 3,850 and 100 DKK, for an expected value of 475 DKK. Thus the two lotteries have a relatively large difference in expected values, in this case 1,165 DKK. As one proceeds down the matrix, the expected value of both lotteries increases, but the expected value of lottery B becomes greater relative to the expected value of lottery A.

In a traditional MPL the subject chooses A or B in each row, and one row is later selected at random for payout for that subject. The logic behind this test for risk aversion is that only risk-loving subjects would take lottery B in the first row, and only very risk-averse subjects would take lottery A in the second last row.[11] Arguably, the last row is simply a test that the subject understood the instructions, and has no relevance for risk aversion at all. A risk neutral subject should switch from choosing A to B when the EV of each is about the same, so a risk-neutral subject would choose A for the first four rows and B thereafter. In our field implementation we instead had the

[11] We are implicitly assuming that the utility function of the subject is only defined over the prizes of the experimental task. We discuss this assumption below.

subject choose on which row to switch from A to B, thus forcing monotonicity, but we also added an option to indicate indifference: we refer to this variant of the MPL as a Sequential MPL (sMPL). For those subjects who did not express indifference we recognized the opportunity to get more refined measures by following up with a subsequent stage where the probabilities attached to the prizes lay within the range of those on the previous switching interval: we refer to this variant as the Iterative MPL (iMPL).[12]

The iMPL uses the same incentive logic as the MPL and sMPL. The logic of selecting a row for payment is maintained but necessitated a revision of the random method used. Let the first stage of the iMPL be called Level 1, the second stage Level 2, and so on. After making all responses, the subject has one row from the first table of responses in Level 1 selected at random by the experimenter. In the MPL that is all there is since there is only a Level 1 table. In the iMPL, that is all there is if the row selected at random by the experimenter is *not* the one at which the subject switched in Level 1. If it *is* the row at which the subject switched, another random draw is made to pick a row in the Level 2 table. For some tasks this procedure is repeated to Level 3.

In order to investigate what effect there may be on responses from using the iMPL we ran lab experiments comparing this procedure to the standard MPL and the sMPL (see Andersen, Harrison, Lau and Rutström [2006]). As noted above, the sMPL changes the MPL to ask the subject to pick the switch point from one lottery to the other, but without the refinement of probabilities allowed in iMPL. Thus it enforces monotonicity, but still allows subjects to express indifference at the "switch" point, akin to a "fat switch point." The subject was then paid in the same manner as

[12] That is, if someone decides at some stage to switch from option A to option B between probability 0.4 and 0.5, the next stage of an iMPL would then prompt the subject to make more choices within this interval for probabilities form 0.40 to 0.50 increasing by 0.01 on each row. The computer implementation of the iMPL restricts the number of stages to ensure that the intervals exceed some *a priori* cognitive threshold (e.g., probability increments of 0.01).

with MPL, but with the non-switch choices filled in automatically.

We used four separate risk aversion tasks with each subject, each with different prizes designed so that all 16 prizes span the range of income over which we seek to estimate risk aversion. The four sets of prizes were as follows, with the two prizes for lottery A listed first and the two prizes for lottery B listed next: (A1: 2000 DKK, 1600 DKK; B1: 3850 DKK, 100 DKK), (A2: 2250 DKK, 1500 DKK; B2: 4000 DKK, 500 DKK), (A3: 2000 DKK, 1750 DKK; B3: 4000 DKK, 150 DKK), and (A4: 2500 DKK, 1000 DKK; B4: 4500 DKK, 50 DKK). At the time of the experiments, the exchange rate was approximately 6.55 DKK per U.S. dollar, so these prizes ranged from approximately $7.65 to $687.

We ask the subject to respond to all four risk aversion tasks and then randomly decide which task and row to play out. In addition, the large incentives and budget constraints precluded paying all subjects, so each subject is given a 10 percent chance to actually receive the payment associated with his decision.

We take each of the binary choices of the subject as the data, and estimate the parameters of a latent utility function that explains those choices using an appropriate error structure to account for the panel nature of the data. Once the utility function is defined, for a candidate value of the parameters of that function, we can construct the expected utility of the two gambles, and then use a linking function to infer the likelihood of the observed choice. We discuss statistical specifications in more detail below.

The MPL instrument has an apparent weakness because it might suggest a frame that encourages subjects to select the middle row, contrary to their unframed risk preferences. The antidote for this potential problem is to devise various "skewed" frames in which the middle row implies different risk attitudes, and see if there are differences across frames. Simple procedures to detect such framing effects, and correct for them statistically if present, have been developed, and

are discussed below (e.g., Harrison, Lau, Rutström and Sullivan [2005], Andersen, Harrison, Lau and Rutström [2006] and Harrison, List and Towe [2007]).

In summary, the set of MPL instruments provides a relatively transparent procedure to elicit risk attitudes. Subjects rarely get confused about the incentives to respond truthfully, particularly when the randomizing devices are physical die that they know that they will toss themselves.[13] As we demonstrate later, it is also possible to infer a risk attitude interval for the specific subject, at least under some reasonable assumptions, as well as to use the choice data to estimate structural parameters of choice models.

C. Estimation Procedures

Two broad methods of estimating risk attitudes have been used. One involves the calculation of bounds implied by observed choices, typically using utility functions which only have a single parameter to be inferred. A major limitation of this approach is that it restricts the analyst to utility functions that can characterize risk attitudes using one parameter. This is because one must infer the bounds that make the subject indifferent between the switch points, and such inferences become virtually incoherent statistically when there are two or more parameters. Of course, for popular functions such as CRRA or Constant Absolute Risk Aversion (CARA) this is not an issue, but if one wants to move beyond those functions then there are problems. It is possible to devise one-parameter functional forms with more flexibility than CRRA or CARA in some dimension, as illustrated nicely by the one-parameter Expo-Power function developed by Abdellaoui, Barrios and Wakker [2007; §4]. But in general we need to move to structural modeling with maximum likelihood

[13] In our experience subjects are suspicious of randomization generated by computers. Given the propensity of many experimenters in other disciplines to engage in deception, we avoid computer randomization whenever feasible.

to accommodate richer models.

The other broad approach involves the direct estimation by maximum likelihood of some structural model of a latent choice process in which the core parameters defining risk attitudes can be estimated, in the manner pioneered by Camerer and Ho [1994; §6.1] and Hey and Orme [1994]. This structural approach is particularly attractive for non-EUT specifications, where several core parameters combine to characterize risk attitudes. For example, one cannot characterize risk attitudes under Prospect Theory without making some statement about loss aversion and probability weighting, along with the curvature of the utility function. Thus joint estimation of all parameters is a necessity for reliable statements about risk attitudes in such cases.[14]

Assume for the moment that utility of income is defined by

$$U(x) = x^{(1-r)}/(1-r) \quad (1)$$

where x is the lottery prize and $r \neq 1$ is a parameter to be estimated. For $r=1$ assume $U(x)=\ln(x)$ if needed. We come back later to the controversial issue of what "x" might be, but for now we assume that it is just the prize shown in the lottery. Thus r is the coefficient of CRRA: $r=0$ corresponds to risk neutrality, $r<0$ to risk loving, and $r>0$ to risk aversion. Let there be two possible outcomes in a lottery. Under EUT the probabilities for each outcome M_j, $p(M_j)$, are those that are induced by the experimenter, so expected utility is simply the probability weighted utility of each outcome in each lottery i plus some level of background consumption ω:

$$EU_i = \sum_{j=1,2} [\, p(M_j) \times U(\omega + M_j) \,]. \quad (2)$$

The EU for each lottery pair is calculated for a candidate estimate of r, and the index

$$\nabla EU = EU_R - EU_L \quad (3)$$

calculated, where EU_L is the "left" lottery and EU_R is the "right" lottery as presented to subjects.

[14] In an important respect joint estimation can be viewed as Full Information Maximum Likelihood (FIML) since it uses the entire set of structural equations from theory to define the overall likelihood.

This latent index, based on latent preferences, is then linked to observed choices using a standard cumulative normal distribution function $\Phi(\nabla EU)$. This "probit" function takes any argument between $\pm\infty$ and transforms it into a number between 0 and 1 using the function shown in Figure 2. Thus we have the probit link function,

$$\text{prob(choose lottery R)} = \Phi(\nabla EU) \tag{4}$$

The logistic function is very similar, as illustrated in Figure 2, and leads instead to the "logit" specification.

Even though Figure 2 is common in econometrics texts, it is worth noting explicitly and understanding. It forms the critical statistical link between observed binary choices, the latent structure generating the index $y^* = \nabla EU$, and the probability of that index y^* being observed. In our applications y^* refers to some function, such as (3), of the EU of two lotteries; or, if one is estimating a Prospect Theory model, the prospective utility of two lotteries. The index defined by (3) is linked to the observed choices by specifying that the R lottery is chosen when $\Phi(\nabla EU)>½$, which is implied by (4).

Thus the likelihood of the observed responses, conditional on the EUT and CRRA specifications being true, depends on the estimates of r given the above statistical specification and the observed choices. The "statistical specification" here includes assuming some functional form for the cumulative density function (CDF), such as one of the two shown in Figure 2. If we ignore responses that reflect indifference for the moment the conditional log-likelihood would be

$$\ln L(\mathbf{r}; y, \omega, \mathbf{X}) = \sum_i [\, (\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1)) \,] \tag{5}$$

where $\mathbf{I}(\cdot)$ is the indicator function, $y_i = 1 (-1)$ denotes the choice of the Option B (A) lottery in risk aversion task i, and $\mathbf{X}$ is a vector of individual characteristics reflecting age, sex, race, and so on. The parameter r is defined as a linear function of the characteristics in vector $\mathbf{X}$.

In most experiments the subjects are told at the outset that any expression of indifference

would mean that if that choice was selected to be played out a fair coin would be tossed to make the decision for them. Hence one can modify the likelihood to take these responses into account by recognizing that such choices implied a 50:50 mixture of the likelihood of choosing either lottery:

$$\ln L(r; y, \omega, \mathbf{X}) = \sum_i [(\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1)) +$$
$$((½ \ln \Phi(\nabla EU) + ½ \ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = 0))] \quad (5')$$

where $y_i = 0$ denotes the choice of indifference. In our experience very few subjects choose the indifference option, but this formal statistical extension accommodates those responses.[15]

The latent index (3) could have been written in a ratio form:

$$\nabla EU = EU_R / (EU_R + EU_L) \quad (3')$$

and then the latent index would already be in the form of a probability between 0 and 1, so we would not need to take the probit or logit transformation. This specification has also been used, with some modifications we discuss later, in Holt and Laury [2002].

Harrison and Rutström [2008; Appendix F] review procedures and syntax from the popular statistical package *Stata* that can be used to estimate structural models of this kind, as well as more complex non-EUT models. The goal is to illustrate how experimental economists can write explicit maximum likelihood (ML) routines that are specific to different structural choice models. It is a simple matter to correct for stratified survey responses, multiple responses from the same subject ("clustering"),[16] or heteroskedasticity, as needed.

---

[15] Our treatment of indifferent responses uses the specification developed by Papke and Wooldridge [1996; equation 5, p.621] for fractional dependant variables. Alternatively, one could follow Hey and Orme [1994; p.1302] and introduce a new parameter τ to capture the idea that certain subjects state indifference when the latent index showing how much they prefer one lottery over another falls below some threshold τ in absolute value. This is a natural assumption to make, particularly for the experiments they ran in which the subjects were told that expressions of indifference would be resolved by the experimenter, but not told how the experimenter would do that (p.1295, footnote 4). It adds one more parameter to estimate, but for good cause.

[16] Clustering commonly arises in national field surveys from the fact that physically proximate households are often sampled to save time and money, but it can also arise from more homely sampling procedures. For example, Williams [2000; p.645] notes that it could arise from dental studies that "collect data

Using the CRRA utility function and equations (1) through (4), we estimate r to be 0.78 for the Danish population, with a standard error of 0.052 and a 95% confidence interval between 0.68 and 0.88. This reflects modest risk aversion over these stakes, and is significantly different from risk-neutrality (r=0).

Extensions of the basic model are easy to implement, and this is the major attraction of the structural estimation approach. For example, one can easily extend the functional forms of utility to allow for varying degrees of relative risk aversion (RRA). Consider, as one important example, the Expo-Power (EP) utility function proposed by Saha [1993]. Following Holt and Laury [2002], the EP function is defined as

$$U(x) = [1 - \exp(-\alpha x^{1-r})]/\alpha, \qquad (1')$$

where α and r are parameters to be estimated. RRA is then $r + \alpha(1-r)y^{1-r}$, so RRA varies with income if $\alpha \neq 0$. This function nests CRRA (as $\alpha \to 0$) and CARA (as $r \to 0$). We illustrate the use of this EP specification in Harrison, Lau and Rutström [2007].

It is also simple matter to generalize this ML analysis to allow the core parameter r to be a linear function of observable characteristics of the individual or task. We would then extend the model to be $r = r_0 + R \times X$, where r_0 is a fixed parameter and R is a vector of effects associated with each characteristic in the variable vector X. In effect the unconditional model assumes $r = r_0$ and just estimates r_0. This extension significantly enhances the attraction of structural ML estimation, particularly for responses pooled over different subjects, since one can condition estimates on

---

on each tooth surface for each of several teeth from a set of patients" or "repeated measurements or recurrent events observed on the same person." The procedures for allowing for clustering allow heteroskedasticity between and within clusters, as well as autocorrelation within clusters. They are closely related to the "generalized estimating equations" approach to panel estimation in epidemiology (see Liang and Zeger [1986]), and generalize the "robust standard errors" approach popular in econometrics (see Rogers [1993]). Wooldridge [2003] reviews some issues in the use of clustering for panel effects, noting that significant inferential problems may arise with small numbers of panels.

observable characteristics of the task or subject.

An important extension of the core model is to allow for subjects to make some errors. The notion of error is one that has already been encountered in the form of the statistical assumption that the probability of choosing a lottery is not 1 when the EU of that lottery exceeds the EU of the other lottery. This assumption is clear in the use of a link function between the latent index ∇EU and the probability of picking one or other lottery; in the case of the normal CDF, this link function is $\Phi(\nabla EU)$ and is displayed in Figure 2. If there were no errors from the perspective of EUT, this function would be a step function, which is shown in Figure 3: zero for all values of $y^*<0$, anywhere between 0 and 1 for $y^*=0$, and 1 for all values of $y^*>0$.

The problem with the CDF of the Hardnose Theorist is immediate: it predicts with probability one or zero. The likelihood approach asks the model to state the probability of observing the actual choice, conditional on some trial values of the parameters of the theory. Maximum likelihood then locates those parameters that generate the highest probability of observing the data. For binary choice tasks, and independent observations, we know that the likelihood of the sample is just the product of the likelihood of each choice conditional on the model and the parameters assumed, and that the likelihood of each choice is just the probability of that choice. So if we have any choice that has zero probability, and it might be literally 1-in-a-million choices, the likelihood for that observation is not defined. Even if we set the probability of the choice to some arbitrarily small, positive value, the log-likelihood zooms off to minus infinity. We can reject the theory without even firing up any statistical package.

Of course, this implication is true for any theory that predicts deterministically, including Expected Utility Theory. This is why one needs some formal statement about how the deterministic prediction of the theory translates into a probability of observing one choice or the other, and then

perhaps also some formal statement about the role that structural errors might play.[17] In short, one *cannot divorce the job of the theorist from the job of the econometrician*, and some assumption about the process linking latent preferences and observed choices is needed. That assumption might be about the mathematical form of the link, as in (1), but it cannot be avoided. Even the very definition of risk aversion needs to be specified using stochastic terms unless we are to impose absurd economic properties on estimates (Wilcox [2008][2010]).

By varying the shape of the link function in Figure 2, one can informally imagine subjects that are more sensitive to a given difference in the index ∇EU and subjects that are not so sensitive. Of course, such informal intuition is not strictly valid, since we can choose any scaling of utility for a given subject, but it is suggestive of the motivation for allowing for structural errors, and why we might want them to vary across subjects or task domains.

Consider the structural error specification used by Holt and Laury [2002], originally due to Luce. The EU for each lottery pair is calculated for candidate estimates of r, as explained above, and the ratio

$$\nabla EU = EU_R^{1/\mu} / (EU_L^{1/\mu} + EU_R^{1/\mu}) \qquad (3'')$$

calculated, where μ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model. The index ∇EU is in the form of a cumulative probability distribution function defined over differences in the EU of the two lotteries and the noise parameter μ. Thus, as $\mu \to 0$ this specification collapses to the deterministic choice EUT model, where the choice is strictly determined by the EU of the two lotteries; but as μ gets larger and larger the choice

[17] Exactly the same insight in a strategic context leads one from Nash Equilibria to Quantal Response Equilibria, if one re-interprets Figures 2 and 3, respectively, in terms of best-response functions defined over expected (utility) payoffs from two strategies. The only difference in the maximum likelihood specification is that the equilibrium condition jointly constrains the likelihood of observing certain choices by two or more players.

essentially becomes random. When $\mu=1$ this specification collapses to (3'), where the probability of picking one lottery is given by the ratio of the EU of one lottery to the sum of the EU of both lotteries. Thus μ can be viewed as a parameter that flattens out the link functions in Figure 2 as it gets larger. This is just one of several different types of error story that could be used, and Wilcox [2008] provides a masterful review of the implications of the alternatives.[18]

There is one other important error specification, due originally to Fechner and popularized by Hey and Orme [1994]. This error specification posits the latent index

$$\nabla EU = (EU_R - EU_L)/\mu \qquad (3''')$$

instead of (3), (3') or (3″).

Wilcox [2008] notes that as an analytical matter the evidence of IRRA in Holt and Laury [2002] would be weaker, or perhaps even absent, if one had used a Fechner error specification instead of a Luce error specification. This important claim, that the evidence for IRRA may be an artefact of the (more or less arbitrary) stochastic identifying restriction assumed, can be tested with the original data from Holt and Laury [2002] and is correct: see Harrison and Rutström [2008; Figure 9].

An important contribution to the characterization of behavioral errors is the "contextual error" specification proposed by Wilcox [2010]. It is designed to allow robust inferences about the primitive "more stochastically risk averse than," and avoids the type of "residual-tail-wagging-the-dog" results that one gets when using the Fechner or Luce specification and the Holt and Laury [2002] data. It posits the latent index

$$\nabla EU = ((EU_R - EU_L)^{\nu})/\mu \qquad (3^*)$$

[18] Some specifications place the error at the final choice between one lottery or after the subject has decided which one has the higher expected utility; some place the error earlier, on the comparison of preferences leading to the choice; and some place the error even earlier, on the determination of the expected utility of each lottery.

instead of (3‴), or

$$\nabla EU = (EU_R/\nu)^{1/\mu} / ((EU_L/\nu)^{1/\mu} + (EU_R/\nu)^{1/\mu}) \qquad (3^{**})$$

instead of (3″), where ν is a new, normalizing term for each lottery pair L and R. The normalizing term ν is defined as the maximum utility over all prizes in this lottery pair minus the minimum utility over all prizes in this lottery pair. The value of ν varies, in principle, from lottery choice to lottery choice: hence it is said to be "contextual." For the Fechner specification, dividing by ν ensures that the *normalized* EU difference $[(EU_R - EU_L)/\nu]$ remains in the unit interval.

D. Asset Integration

There is a tension between experimental economists and theorists over the proper interpretation of estimates of risk attitudes that emerge from experimental choice behavior. Experimental economists claim to provide evidence of risk aversion over small stakes, which we take here to be amounts such as $10, $100 or even several hundred dollars. Some theorists argue that these estimates are "implausible," in a sense to be made explicit. Although the original arguments of theorists were couched as attacks on the plausibility of EUT (e.g., Hansson [1988] and Rabin [2002]), it is now apparent that the issues are just as important, or unimportant, for non-EUT models (e.g., Safra and Segal [2008] and Cox and Sadiraj [2008]).

The notion of plausibility can be understood best by thinking of the argument in several steps, even if they are often collapsed into one. First, someone proposes *point* estimates of some utility function. These estimates typically come from inferences based on observed choice behavior in an experiment, derived from maximum likelihood estimates of a structural model of latent choice behavior (e.g., Harrison and Rutström [2008; §3]). Standard errors on those estimates are usually not relied on in these exercises. Second, with some auxiliary assumptions, an analyst constructs a lottery choice task in which these point estimates generate predictions on some domain of lottery prizes,

typically involving at least one prize that is much larger than the domain of prizes over which the estimates were derived (e.g., Rabin [2002], Cox and Sadiraj [2008; §3.2]). In some cases these predictions are also defined on the domain of lottery prizes of the experimental tasks (e.g., Cox and Sadiraj [2008; §4.5]). Third, the analyst views this constructed lottery choice task as a "thought experiment," in the sense that it is just like an actual experiment except that it is not actually implemented (Harrison and List [2004; §9]). One reason not to conduct the experiment is that it might involve astronomic stakes, but the main reason is that it is assumed *a priori* obvious what the choice would be, in some sense eliminating the need for additional experiments. Finally, it is pointed out that the predicted outcome from the initial estimates is contrary to the *a priori* obvious choice in the thought experiment. Thumbs down, and the initial estimates are discarded as implausible.

Since the initial estimates are typically defined over observed experimental choices with small stakes, we refer here to the implied claims about risk aversion as "risk aversion in the small." The predicted behavior in the thought experiment is typically defined over choices with very large stakes, so we refer to the implied choices as reflecting "risk aversion in the large." So plausibility can be viewed as a tension and inconsistency between observations (real and imagined *a priori*) generated on two domains. The issue is not that the subject has to have the same relative or absolute measure of risk aversion for different prizes: these problems arise even when "flexible" functional forms are employed for utility functions.

One general response is to just focus on risk attitudes in the small, and make no claims about behavior beyond the domain over which the estimates were obtained. This position states that if one had estimated over larger domains, then the estimated models would reflect actual choices over that domain, but one simply cannot apply the risk aversion estimates outside the domain of estimation. Since large parts of economic theory are written in terms of the utility of income, rather than the utility of wealth, this approach has some validity. Of course, there is nothing in principle to stop

one defining income as a large number, either with a large budget, subjects in a very poor country (e.g., Harrison, Humphrey and Verschoor [2010]), or by using natural experiments such as game shows (e.g., Andersen, Harrison, Lau and Rutström [2007b]).

A second approach, which we employed in Andersen, Harrison, Lau and Rutström [2008a], was to assume some level of baseline consumption that was suggested by expenditure data for the subjects.

A third approach is to test for the degree of asset integration in observed behavior. If one adopts a general specification, following Cox and Sadiraj [2006], and allows income and wealth to be arguments of some utility function, then one does not have to assume that the argument of the utility function is income or wealth. One might posit an aggregation function that combines the two in some way, and this composite then being evaluated with some standard utility function. For example, assume the linear aggregation function $\omega W + y$, where W is wealth, y is experimental income, and ω is some weighting parameter to be assumed or estimated. Or one could treat ωW and y as inputs into some Constant Elasticity of Substitution function, and estimate or assume ω and the elasticity of substitution. This approach allows the popular special cases of zero asset integration and perfect asset integration, but lets the "data decide" when these parameters are estimated in the presence of actual choices. Where does one get estimates of W? As it happens, very good proxies for W can be inferred from data in Denmark that is collected by *Statistics Denmark*, the official government statistics agency. One can then calculate those proxies for subjects that have been in experiments such as ours (such things being feasible in Denmark, with appropriate confidentiality agreements), and estimate the weighting parameter. Preliminary estimates suggest that ω is very small indeed, and that the elasticity of substitution between ωW and y is close to 1.

3. Discount Rates

In many settings in experimental economics we want to elicit some preference from a set of choices that also depend on risk attitudes. Often these involve strategic games, where the uncertain ways in which behavior of others deviate from standard predictions engenders a lottery for each player. Such uncertain deviations could be due to, for example, unobservable social preferences such as fairness or reciprocity. One example is the offer observed in Ultimatum bargaining when the other player cannot be assumed to always accept a minuscule amount of money, and acceptable thresholds may be uncertain. Other examples include Public goods contribution games where one does not know the extent of free riding of other players, Trust games in which one does not know the likelihood that the other player will return some of the pie transferred to him, or Centipede games where one does not know when the other player will stop the game. Another source of uncertainty is the possibility that subjects make decisions with error, as predicted in Quantal Response Equilibria. Harrison [1987] and Harrison and Rutström [2008; §3.6] consider the use of controls for risk attitudes in bidding in first-price auctions.

In some cases, however, we simply want to elicit a preference from choices that do not depend on the choices made by others in a strategic sense, but which still depend on risk attitudes in a certain sense. An example due to Andersen, Harrison, Lau and Rutström [2008a] is the elicitation of individual discount rates. In this case it is the concavity of the utility function that is important, and under EUT that is synonymous with risk attitudes. Thus the risk aversion task is just a (convenient) vehicle to infer utility over deterministic outcomes. The implication is that we should combine a risk elicitation task with a time preference elicitation task, and use them jointly to infer discount rates over utility.

A. Defining Discount Rates in Terms of Utility

Assume EUT holds for choices over risky alternatives and that discounting is exponential. A subject is indifferent between two income options M_t and $M_{t+\tau}$ if and only if

$$U(\omega+M_t) + (1/(1+\delta)^{\tau})\, U(\omega) = U(\omega) + (1/(1+\delta)^{\tau})\, U(\omega+M_{t+\tau}) \qquad (6)$$

where $U(\omega+M_t)$ is the utility of monetary outcome M_t for delivery at time t plus some measure of background consumption ω, δ is the discount rate, τ is the horizon for delivery of the later monetary outcome at time $t+\tau$, and the utility function U is separable and stationary over time. The left hand side of equation (6) is the sum of the discounted utilities of receiving the monetary outcome M_t at time t (in addition to background consumption) and receiving nothing extra at time $t+\tau$, and the right hand side is the sum of the discounted utilities of receiving nothing over background consumption at time t and the outcome $M_{t+\tau}$ (plus background consumption) at time $t+\tau$. Thus (6) is an indifference condition and δ is the discount rate that equalizes the present value of the *utility* of the two monetary outcomes M_t and $M_{t+\tau}$, after integration with an appropriate level of background consumption ω.

Most analyses of discounting models implicitly assume that the individual is risk neutral,[19] so that (6) is instead written in the more familiar form

$$M_t = (1/(1+\delta)^{\tau})\, M_{t+\tau} \qquad (7)$$

where δ is the discount rate that makes the present value of the two monetary outcomes M_t and $M_{t+\tau}$ equal.

To state the obvious, (6) and (7) are not the same. As one relaxes the assumption that the decision maker is risk neutral, it is apparent from Jensen's Inequality that the implied discount rate

[19] See Keller and Strazzera [2002; p. 148] and Frederick, Loewenstein and O'Donoghue [2002; p.381ff.] for an explicit statement of this assumption, which is often implicit in applied work. We refer to risk aversion and concavity of the utility function interchangeably, but it is concavity that is central (the two can differ for non-EUT specifications).

decreases if U(M) is concave in M. Thus one cannot infer the level of the individual discount rate without knowing or assuming something about their risk attitudes. This identification problem implies that risk attitudes and discount rates cannot be estimated based on discount rate experiments alone, but separate tasks to identify the influence of risk preferences must also be implemented.

Thus there is a clear implication from theory to experimental design: you need to know the non-linearity of the utility function before you can *conceptually* define the discount rate. There is also a clear implication for econometric method: you need to jointly estimate the parameters of the utility function and the discount rate, to ensure that sampling errors in one propagate correctly to sampling errors of the other. In other words, if we know the parameters of the utility function less precisely, due to small samples or poor parametric specifications, we have to use methods that reflect the effect of that imprecision on our estimates of discount rates.

Andersen, Harrison, Lau and Rutström [2008a] do this, and infer discount rates for the adult Danish population that are well below those estimated in the previous literature that assumed risk neutrality, such as Harrison, Lau and Williams [2002], who estimated annualized rates of 28.1% for the same target population. Allowing for concave utility, they obtain a point estimate of the discount rate of 10.1%, which is significantly lower than the estimate of 25.2% for the same sample assuming linear utility. This does more than simply verify that discount rates and risk aversion coefficients are mathematical substitutes in the sense that either of them have the effect of lowering the influence from future payoffs on present utility. It tells us that, for risk aversion coefficients that are reasonable from the standpoint of explaining choices in the lottery choice task, the estimated discount rate takes on a value that is much more in line with what one would expect from market interest rates. To evaluate the statistical significance of adjusting for a concave utility function one can test the hypothesis that the estimated discount rate assuming risk aversion is the same as the discount rate estimated assuming risk neutrality. This null hypothesis is easily rejected. Thus, *allowing*

for risk aversion makes a significant difference to the elicited discount rates.

B. The Need for Joint Estimation

We can write out the likelihood function for the choices that our subjects made and jointly estimate the risk parameter r in equation (1) and the discount rate δ. We use the same stochastic error specification as Holt and Laury [2002], and the contribution to the overall likelihood from the risk aversion responses is given by (5').

A similar specification is employed for the discount rate choices. Equation (3) is replaced by the discounted utility of each of the two options, conditional on some assumed discount rate, and equation (4) is defined in terms of those discounted utilities instead of the expected utilities. The discounted utility of Option A is given by

$$PV_A = (\omega+M_A)^{(1-r)} + (1/(1+\delta)^{\tau})\ \omega^{(1-r)} \quad (8)$$

and the discounted utility of Option B is

$$PV_B = \omega^{(1-r)} + (1/(1+\delta)^{\tau})\ (\omega+M_B)^{(1-r)} \quad (9)$$

where M_A and M_B are the monetary amounts in the choice tasks presented to subjects, illustrated in Table 2, and the utility function is assumed to be stationary over time.

An index of the difference between these present values, conditional on r and δ, can then be defined as

$$\nabla PV = PV_B^{1/\eta} / (PV_A^{1/\eta} + PV_B^{1/\eta}) \quad (10)$$

where η is a noise parameter for the discount rate choices, just as μ was a noise parameter for the risk aversion choices. It is not obvious that $\mu=\eta$, since these are cognitively different tasks. Our own priors are that the risk aversion tasks are harder, since they involve four outcomes compared to two outcomes in the discount rate tasks, so we would expect $\mu>\eta$. Error structures are things one should always be agnostic about since they capture one's modeling ignorance, and we allow the error

terms to differ between the risk and discount rate tasks.

Thus the likelihood of the discount rate responses, conditional on the EUT, CRRA and exponential discounting specifications being true, depend on the estimates of r, δ, μ and η, given the assumed value of ω and the observed choices.[20] If we ignore the responses that reflect indifference, the conditional log-likelihood is

$$\ln L(r, \delta, \mu, \eta; y, \omega, \mathbf{X}) = \sum_i [(\ln \Phi(\nabla PV) \times \mathbf{I}(y_i=1)) + (\ln (1-\Phi(\nabla PV)) \times \mathbf{I}(y_i=-1))] \quad (11)$$

where $y_i = 1(-1)$ again denotes the choice of Option B (A) in discount rate task i, and $\mathbf{X}$ is a vector of individual characteristics.

The joint likelihood of the risk aversion and discount rate responses can then be written as

$$\ln L(r, \delta, \mu, \eta; y, \omega, \mathbf{X}) = \ln L^{RA} + \ln L^{DR} \quad (12)$$

where L^{RA} is defined by (5′) and L^{DR} is defined by (11). This expression can then be maximized using standard numerical methods.

4. Lessons Learned

We draw together some methodological and practical lessons we have learned from our research into risk and time preferences in Denmark. These are often reflections on our experience over many years in considering how theory, experimental design and econometrics inform and constrain each other.

[20] For simplicity we are implicitly assuming that the λ parameter from Andersen, Harrison, Lau and Rutström [2008a] is equal to 1. This means that delayed experimental income is spent in one day.

A. The Role of Artefactual Field Experiments

Harrison and List [2004] go to great lengths to point out that field experiments often entail many changes compared to traditional laboratory experiments. Sample composition, type of commodity, environment, information, stakes, literacy, and so on. When one observes differences in behavior in the field compared to the lab, which of these is driving that difference? Or, how do we know that there are not offsetting effects from different components of the field environment? For some inferential purposes we don't care about this sort of decomposition, but all too often we care deeply. The reason is that the "story" or "spin" that is put on the difference in behavior has to do with some structural component of the theory explaining behavior. For example, the claim in some quarters that people exhibit more rational behavior in the field, and that irrationality is primarily confined to the lab.[21] Or that people exhibit apparent altruism in the lab but rarely in the field. Or that "market interactions" fix all evils of irrationality. These are overstatements, but are not too far from the sub-plot of recent literature.

The critical role of "artefactual field experiments," as Harrison and List [2004] call them, is to take the simplest conceptual step over the bridge from the lab to the field: vary the composition of the sample from the convenience sample of university students. Although conceptually simple, the implementation is not always simple, particularly if one wants to generate representative samples of a large population as distinct from studying well-defined sub-samples that mass conveniently at trade shows or other locations. Whether these are best characterized as being lab experiments or field experiments is not, to us, the real issue: the key thing is to see this type of experiment along a continuum taking one from the lab to the field, to better understand behavior.

To illustrate this path, consider the evaluation of risk attitudes in the field. Our field

[21] And hence that we can safely dismiss the messy claims of behaviorists as artefacts of the lab. We might agree with this conclusion even if we do not agree with this argument for it (Harrison [2010]).

experiments in Denmark, reviewed in section 2, illustrate well the issues involved in taking the first step away from the lab. But to go further entails more than just "leaving the classroom" and recruiting outside of a university setting. In terms of sample composition, it means finding subjects who deal with that type of uncertainty in varying degrees, and trying to measure the extent of their field experience with uncertainty. Moreover, it means developing stimuli that more closely match those that the subjects have previously experienced, so that they can use whatever heuristics they have developed for that commodity when making their choices. Finally, it means developing ways of communicating probabilities that correspond with language that is familiar to the subject. Thus, field experimentation in this case ultimately involves several simultaneous changes from the lab setting with respect to subject recruitment and the development of stimuli that match the field setting. Examples of studies that do these things, to varying degrees, are Harrison, List and Towe [2007] and Fiore, Harrison, Hughes and Rutström [2009]. In each case the changes were, by design, partial, so that one could better understand the effect on behavior. This is the essence of control that leads us to use experimental methods, after all.

To see the concern with going into the field from the lab, consider the importance of "background risk" for the attitudes towards a specific "foreground risk" that are elicited. In many field settings it is simply not possible to artificially identify attitudes towards one risk source without worrying about how the subjects view that risk as being correlated with other risks. For example, mortality risks from alternative occupations tend to be highly correlated with morbidity risks: what doesn't kill you, sadly, often injures you. It is implausible to ask subjects their attitude toward one risk without some coherent explanation in the instructions as to why a higher or lower level of that risk would not be associated with a higher or lower risk of the other. In general this will not be something that is amenable to field investigation in a controlled manner, although a few exceptions exist, as illustrated by Harrison, List and Towe [2007].

In a similar vein, there is a huge literature on how one can use laboratory experiments to calibrate hypothetical field surveys for "hypothetical bias" in valuations: see Harrison [2006] for a review. In this case the value of the complementarity of field and lab is not due to concerns about the artefactual nature of the lab, but it is rather the artefactual nature of the field commodity that is causing problems. For example, when someone asks you your willingness to pay $100 to reduce the risk of global warming, how should you interpret what you are actually buying? There is simply no way to run a naturally-occurring field experiment in this case, or in any way that is free of major confounds. So evidence of hypothetical bias in many, disparate private goods experiments can be used to condition the responses obtained in the field. For example, if women always respond identically in hypothetical and real lab experiments, and men state valuations that are always double what they would if it were real, then one surely has *some* basis for adjusting field hypothetical responses if one knows the sex of the respondent. The notion of calibration, introduced in this area by Blackburn, Harrison and Rutström [1994], formalizes the statistical process of adjusting the field survey responses for the priors that one obtains in the lab environment.

B. The Contrived Debate Between Lab and Field Experiments

A corollary of the case for artefactual field experiments is the case for the complementarity of laboratory and field experiments. This theme was front and center in Harrison and List [2004], but appears to have been lost in some subsequent commentaries selling field experimental methodology. One illustration of this complementarity comes from our work and is motivated by the view that theory, experimental design and econometrics are dependant, resulting in numerous auxiliary hypotheses that can most efficiently be investigated in the lab. It is often unwieldy and inefficient to test procedures and treatments completely in the field, even if one would like to do so. The efficient mix is to identify the important treatments for application in the field, and address the

less important ones in the laboratory. Of course this involves some judgement about what is important, but we are often guided there by theory, previous evidence, the need for econometric identification, and, yes, one's nose for what sells in the journals. We have been careful in our own work to consider the balance between lab and field carefully at the outset. We expect that many others will do the same, so that the tendency to present field experiments as universally superior to lab experiments will organically shrink.

There are several examples of this complementarity from our work. In one case we used the laboratory to evaluate the performance of the iMPL elicitation procedure we assumed in the field. In the lab we could consider controlled comparisons to alternative methods, and see what biases might have been generated by using the iMPL (see Andersen, Harrison, Lau and Rutström [2006a]). It would simply have been inefficient to carry all of those variants into the field.

In another example, we were concerned about the possibility of sample selection into experiments on the basis of risk attitudes: the so-called "randomization bias" much discussed in the broader experimental literature. If subjects know that participation in experiments might entail randomization to treatment, and they have heterogeneous risk attitudes, then one would *a priori* expect to see less risk averse subjects in experiments. But in experimental economics we offset that with a fixed, non-stochastic show-up fee, so what is the net effect? To be honest, we only thought of this after running our previous generation of field experiments, but could quickly evaluate the obvious experimental design in the lab with the same instruments (see Harrison, Lau and Rutström [2009]). Finding evidence of sample selection, we now build the obvious design checks into the next generation of our field experiments.

C. Danes Are Like Plain Yogurt, Not Like Wines Or Cheeses

We are well aware that results for Denmark, as important as they are for Danish policy, and

perhaps also methodologically, might not readily transfer to other populations. In our case much of our non-Danish work has involved developing countries, and the differences there can be dramatic. We would expect to see greater heterogeneity, greater instability, and perhaps even greater variety in the types of decision-making models employed. In this respect, however, we stress that the tools we have developed may be generally applied.

To take one important example, consider what one might mean by the "stability" of risk preferences over time. Does this mean that the unconditional estimate of RRA for each subject is the same over time, that the distribution for a given population stays the same even if individuals pop up at different parts of the distribution from time to time, or that the RRA or distribution are stable functions of observable states of nature that might change over time? In the latter case, where we think of states of nature as homely and intelligible things such as health, marital status, and family composition, it could be that preferences are a stable function of those states, but appear to be unstable when evaluated unconditionally. Using a longitudinal field experimental design, we examined exactly this question in Denmark (Anderson, Harrison, Lau and Rutström [2008b]). We found that preferences were generally stable, with some caveats, in virtually all three senses. But is this just a reflection of the "plain yogurt" of Danish culture, or something more general? Our personal priors may tell us one thing, but only data from new experiments can verify if this is true. But the point is that the longitudinal methodology, and questionnaires on states of nature, is a general approach applicable beyond the specific population of Denmark.

D. Non-EUT Models of Risky Choice and Non-Exponential Models of Discounting: *Festine Lente*

We have serious doubts about the generality and robustness of some of the empirical claims of the literature with respect to risk and time preferences. Much of the empirical evidence has been

obtained by staring at "patterns" of choices rather than estimating structural parameters and testing for statistical significance. It is quite possible for there to be statistically significant differences in patterns of choices, according to some unconditional semi-parametric test, but for that to be consistent with a wide range of underlying structural models. This is particularly true when one augments those models with alternative "behavioral error" stories. To take one simple example, the violations of first-order stochastic dominance that motivated the shift from original prospect theory to cumulative prospect theory can, to some extent, be accounted for by certain error specifications. One might not want to do that, and we are not advocating it as a general econometric policy, but the point is that inferences about patterns is not the same as inferences about the latent structural parameters.

Another issue with much of the received evidence for violations of EUT or exponential discounting is that it has been drawn from convenience samples in laboratory experiments. Relatively little evidence has been collected from field experiments with a broader sample from the population, using comparable instruments and/or instruments that arise more naturally in the decision-making environments of the subjects. We are not denying that "students are people too," just noting that they have distinct set of demographic characteristics that can matter significantly for policy inferences (e.g., Andersen, Harrison, Lau and Rutström [2010]), and that all exhibit one characteristic that might reflect sample selection on unobservables of relevance for the experimental task at hand (viz., their presence at a college or university).

Finally, our work leads us to question some of the inferential assumptions of previous tests. Mixture specifications, in which one allows two or more data-generating processes to explain observed behavior, show clear evidence that behavior is not wholly explained by any one of the popular models. Andersen, Harrison, Lau and Rutström [2008a; §3.D] consider a mixture specification of exponential and hyperbolic discounting, and find that 72% of the choices are better

characterized as exponential.[22] This estimate of the mixing probability is statistically significantly different from 0 or 50%. Similarly, Harrison and Rutström [2009] find roughly equal support for EUT and Prospect Theory in a lab setting; Harrison, Humphrey and Verschoor [2009] find roughly equal support for EUT and Rank-Dependent Utility models in artefactual field experiments in India, Ethiopia and Uganda; and Coller, Harrison and Rutström [2010] find roughly equal support for exponential and quasi-hyperbolic discounting in the laboratory.

The key insight from mixture specifications is to simply change the question that is posed to the data. Previous econometric analyses have posed a proper question: if one and only one data-generating process is to account for these data, what are the estimated parameter values and do they support a non-standard specification? The simplest, finite mixture specification changes this to: if two data-generating processes are allowed to account for the data, what fraction is attributable to each, and what are the estimated parameter values? So stated, one can imagine someone still wanting to ask the former question, if they just wanted one "best" model. But that question is also seen to constrain evidence of heterogeneity of decision-making processes, and we prefer to avoid that when we can. There are fascinating issues with the specific implementation and interpretation of mixture models, but those are not germane to the main insight they provide.[23]

Finally, there are often simple issues of functional specification which can only be explored with structural models. For example, what happens when subjects bring an unobserved

---

[22] Those experiments employed a Front End Delay on payments of 30 days, and were not designed to test Quasi-Hyperbolic specifications. The latest series of field experiments in Denmark, completed late 2009, are designed to test that specification *inter alia*.

[23] For example, does one constrain individuals or task types to be associated with just one data-generating process, or allow each choice to come from either? Does one consider more than two types of processes, using some specification rule to decide if there are 2, or 3, or more? Does one specify general models for each data-generating process and see if one of them collapses to a special case, or just specify the competing alternatives explicitly from the outset? How does one check for global maximum likelihood estimates in an environment that might generate multi-modal likelihood functions "naturally"? Harrison and Rutström [2009] discuss these issues, and point to the older literature.

"homegrown reference point" to the experimental task and the analyst tries to infer measures of loss aversion? In other words, what if the subject rationally expects to get more than just the show-up fee? The answer is that one gets extremely sensitive estimates of loss aversion depending on what one assumes, not too surprisingly (e.g., Harrison and Rutström [2008; §3.2.3]). This is likely to be a more serious issue in the field, due to a greater diversity in homegrown reference points. As another example, some tests of EUT rest on the assumption that subjects use a very restrictive functional form. The tests of myopic loss aversion offered in Gneezy and Potters [1997] rest on assuming CRRA, and their "violations of EUT" can be accounted for simply with an expo-power specification that allows varying RRA with prize level (e.g., Harrison and Rutström [2008; §3.7]). So the initial evidence does show a violation of CRRA, but that is not something one ought to get too excited about.

We do not want to overstate the case for standard specifications. We do find some evidence for *some* subjects to behave differently than typically assumed in EUT and exponential discounting specifications, in *some* tasks. It is just that the evidence is hardly as monolithic as many claim.

E. Estimation, Not Direct Elicitation

When we began our research we were focused on designing instruments that could directly provide more precise estimates of risk attitudes and temporal preferences. Our work on the iMPL, discussed earlier, was directly solely at that objective: refining down the interval within which we had captured the true, latent risk attitude or discount rate. We certainly believe that those procedures accomplish that goal, but our focus quickly moved away from designing a better mousetrap to learning more about the mouse itself. That is, we discovered that the questions we wanted to ask demanded that we employ structural estimation, if for no other reason than to be able to condition inferences from the discount rate task with estimates of the utility function (from the risk aversion

task). This need for joint estimation, and full information maximum likelihood recognition that errors in estimation of the utility function *should* propagate into errors in estimation of discount rates, as a matter of theory as well as econometric logic, is much more general than these examples. Inferences about bidding in auctions, about subjective beliefs elicited from a proper scoring rule, about social preferences, and behavior in games with payoffs defined over utility, all require that one say something about utility functions unless one is working on special cases. Indeed, there is mounting evidence that the inferences change dramatically when one allows for non-linear utility functions.[24] The day of designing tasks to elicit exactly what one wants in one task are long over.

F. Virtual Experiments As A Smooth Bridge Between the Lab and the Field

It is now well-known and accepted that behavior is sensitive to the cognitive constraints of participants. It has been recognized for some time that field referents and cues are essential elements in the decision process, and can serve to overcome such constraints (Ortmann and Gigerenzer [1997]), even if there are many who point to "frames" as the source of misbehavior from the perspective of traditional economic theory (Kahneman and Tversky [2000]). The concept of "ecological rationality" captures the essential idea of those who see heuristics as potentially valuable decision tools (Gigerenzer and Todd [1999], Smith [2003]). According to this view, cognition has evolved within specific decision environments. If that evolution is driven by ecological fitness then the resulting cognitive structures, such as decision heuristics, are efficient and accurate *within these environments*. But they may often fail when applied to new environments.

At least two other research programs develop similar views. Glimcher [2003] describes a research program, following Marr [1982], that argues for understanding human decision making as a

[24] There is no intended slight of models of decision-making under risk that focus on things other than the linearity of the utility function: this point is quite general.

function of a complete biological system rather than as a collection of mechanisms. As a biological system he views decision making functions as having evolved to be fit for specific environments. Clark [1997] sees cognition as extended outside not just the brain but the entire human body, defining it in terms of all the tools used in the cognitive process, both internal and external to the body. Field cues can be considered external aspects of such a process. Behavioral economists are paying attention to these research programs and what they imply for the understanding of the interactions between the decision maker and his environment. For our purposes here, it means we have to pay careful attention to the role of experiential learning in the presence of specific field cues and how this influences decisions.

The acceptance of the role of field cues in cognition provides arguments in favor of field rather than lab experiments (Harrison and List [2004]). Where else than in field experiments can you study decision makers in their natural environment using field cues that they have come to depend on? We actually challenge this view, if it is taken to argue that the laboratory environment is *necessarily* unreliable (Levitt and List [2007]). While it is true that lab experiments traditionally use artefactual and stylized tasks that are free of field cues, in order to generate the type of control that is seen as essential to hypothesis testing, field experiments have other weaknesses that *a priori* are equally important to recognize (Harrison [2005]). Most importantly, the ability to implement necessary controls on experimental conditions in the field is much more limited than in the lab, as is the ability to implement many counterfactual scenarios. In addition, recruitment is often done in such a way that it is difficult to avoid and control for sample selection effects; indeed, in many instances the natural process of selection provides the very treatment of interest (e.g., Harrison and List [2008]). However, that means that one must take the sample with all of the unobservables that it might have selected on, and just assume that they did not interact with the behavior being measured. Finally, the cost of generating observational data can be quite significant in the field, at least in

comparison to the lab.

For all these reasons we again see lab and field experiments as complementary, a persistent theme of Harrison and List [2004]. A proper understanding of decision making requires the use of both. While lab experiments are better at generating internal validity, imposing the controlled conditions necessary for hypothesis testing, field experiments are better at generating external validity, including the natural field cues.

Fiore, Harrison, Hughes and Rutström [2009] propose a new experimental environment, the Virtual Experiment (VX), that has the potential of generating both the internal validity of lab experiments and the external validity of field experiments. A VX is an experiment set in a controlled lab-like environment, using either typical lab or field participants, that generates synthetic field cues using Virtual Reality (VR) technology. The experiment can be taken to typical field samples, such as experts in some decision domain, or to typical lab samples, such as student participants. The VX environment can generate internal validity since it is able to closely mimic explicit and implicit assumptions of theoretical models, and thus provide tight tests of theory; it is also able to replicate conditions in past experiments for robustness tests of auxiliary assumptions or empirically generated hypotheses. The VX environment can generate external validity because observations can be made in an environment with cues mimicking those occurring in the field. In addition, any dynamic scenarios can be presented in a realistic and physically consistent manner, making the interaction seem natural for the participant. Thus the VX builds a bridge between the lab and the field, allowing the researcher to smoothly go from one to the other and see what features of each change behavior. VX is a methodological frontier enabling new levels of understanding via integration of laboratory and field research in ways not previously possible. Echoing calls by others for such an integration, we argue that "research must be conducted in various settings, ranging from the artificial laboratory, through the naturalistic laboratory, to the natural environment itself" (Hoffman and Deffenbacher

[1993; p. 343]).

The potential applications for VX are numerous. Apart from simulating actual policy scenarios, such as the wild fire prevention policies investigated by Fiore et al. [2009], it can also be used to mimic environments assumed in a number of field data analyses. For example, popular ways of estimating valuations for environmental goods include the Travel Cost Method (TCM), the Hedonic Pricing Method (HPM), and the Stated Choice Method (SCM). To mimic TCM the simulation can present participants with different travel alternatives and observe which ones are chosen under different naturalistic conditions. To mimic HPM the simulation can present participants with different real estate options and observe purchasing behavior, or simply observe pricing behavior for alternative options (Castronova [2004]). Finally, to mimic SCM participants can experience the different options they are to choose from through naturalistic simulation. For all of these types of scenarios, some of the most powerful applications of VX will involve continuous representations of dynamically generated effects of policy changes. Visualizing and experiencing long-term effects correctly should improve short-run decisions with long-run consequences.

In the application to wild fire prevention policies we use actual choices by subjects that bear real economic consequences from those choices. Participants are presented with two options: one simply continues the present fire prevention policies, and the other increases the use of prescribed burns. Participants get to experience two fire seasons under each policy and are then asked to make a choice between them. The scenario that simulates the continuation of the present fire prevention policies will realistically generate fires that cause more damage on average and that also vary substantially in intensity. This option therefore presents the participant with a risky gamble with low expected value. The alternative option presents a relatively safe gamble with a higher expected value, but there will be a non-stochastic cost involved in implementing the expansion of prescribed burns. It is possible in VX to set the payoff parameters in such a way that one can estimate Willingness To

Pay (WTP) for the burn expansion option that is informative to actual fire policy. These values of WTP could then be compared to those generated through a popular Contingent Valuation Method to test the hypothesis that they should be different. Alternatively, it is possible to manipulate the payoff parameters in such a way that one estimates parameters of choice models such as risk attitudes, loss aversion, and probability weights.

In summary, we see the use of VX as a new tool in experimental economics, with an emphasis on the methodological issues involved in bridging the gap between lab and field experiments.

5. Conclusions

We love doing experimental economics because it provides a unique ability to directly and explicitly confront theorists and econometricians with real behavior. There is no hiding behind theoretical models that claim to be operationally meaningful but never really come close, and there is no hiding behind proxies for variables of theoretical interest when the experimental design is developed rigorously. This work requires modest extensions of the tools used by experimental economists, to develop designs that employ several tasks to identify all of the moving parts of the theoretical machine, and to write out the likelihood of observed behavior using the structural model being tested, but these are feasible and well-understood components of modern experimental economics.

It is then natural to combine laboratory and field experiments in this enterprise. When we think of new areas of our own research, such as the elicitation of subjective beliefs and the measurement of aversion to uncertainty, we would never think of first developing experimental designs in the field. Actually, this is also true of older areas of research which we believe to be often intractably confounded in the field, such as "social preferences," "trust" and "loss aversion," but our

premiss there is apparently not widely shared. In any event, the efficient frontier for this production process demands that we combine laboratory and field environments. They can be substituted to varying degrees, of course, but it is difficult for us to imagine an area of enquiry that would not benefit from some inputs of both.

Table 1: Typical Payoff Matrix in the Danish Risk Aversion Experiments

Lottery A				Lottery B				EV^A	EV^B	Difference	Open CRRA Interval
p	DKK	p	DKK	p	DKK	p	DKK	DKK	DKK	DKK	if Subject Switches to Lottery B and ω=0
0.1	2000	0.9	1600	0.1	3850	0.9	100	1640	475	1165	-∞, -1.71
0.2	2000	0.8	1600	0.2	3850	0.8	100	1680	850	830	-1.71, -0.95
0.3	2000	0.7	1600	0.3	3850	0.7	100	1720	1225	495	-0.95, -0.49
0.4	2000	0.6	1600	0.4	3850	0.6	100	1760	1600	160	-0.49, -0.15
0.5	2000	0.5	1600	0.5	3850	0.5	100	1800	1975	-175	-0.15, 0.14
0.6	2000	0.4	1600	0.6	3850	0.4	100	1840	2350	-510	0.14, 0.41
0.7	2000	0.3	1600	0.7	3850	0.3	100	1880	2725	-845	0.41, 0.68
0.8	2000	0.2	1600	0.8	3850	0.2	100	1920	3100	-1180	0.68, 0.97
0.9	2000	0.1	1600	0.9	3850	0.1	100	1960	3475	-1515	0.97, 1.37
1	2000	0	1600	1	3850	0	100	2000	3850	-1850	1.37, ∞

Note: The last four columns in this table, showing the expected values of the lotteries and the implied CRRA intervals, were not shown to subjects.

Figure 1: An Illustrative Policy Lottery

Distribution of welfare effects of indirect tax uniformity

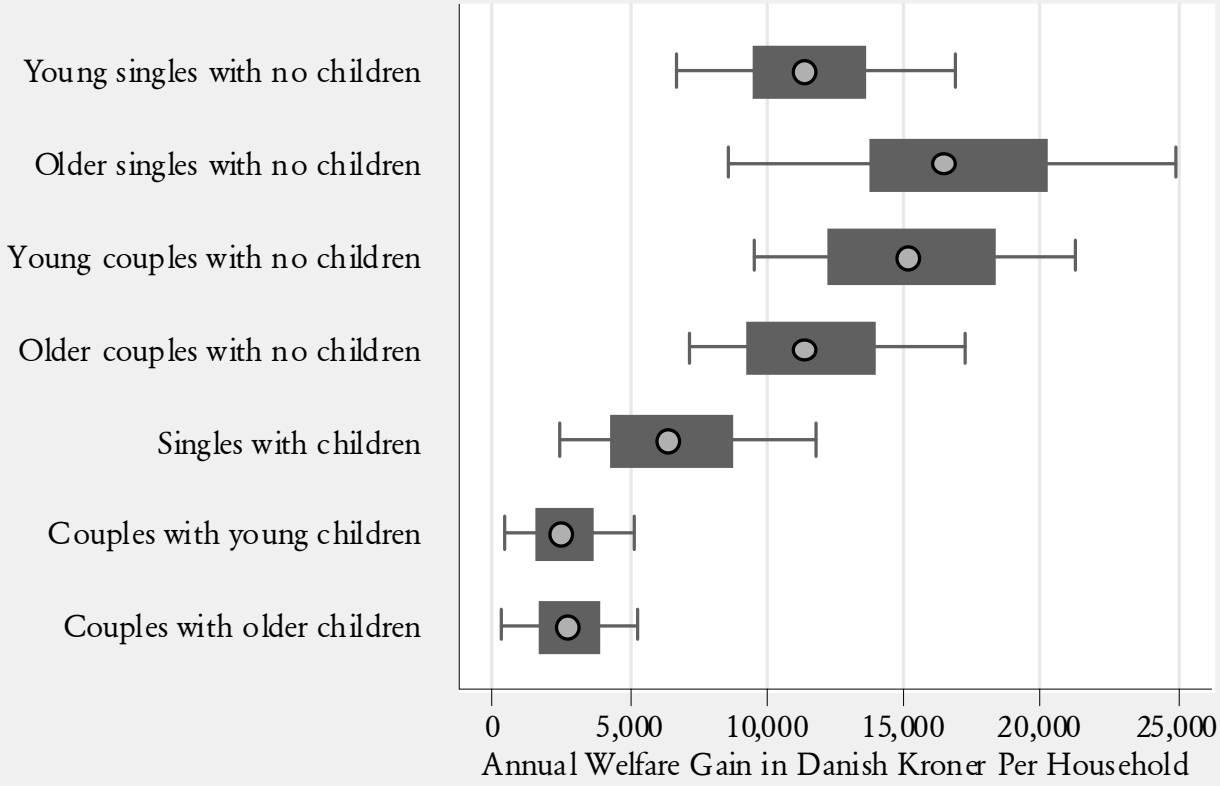


Figure 2: Normal and Logistic Cumulative Density Functions

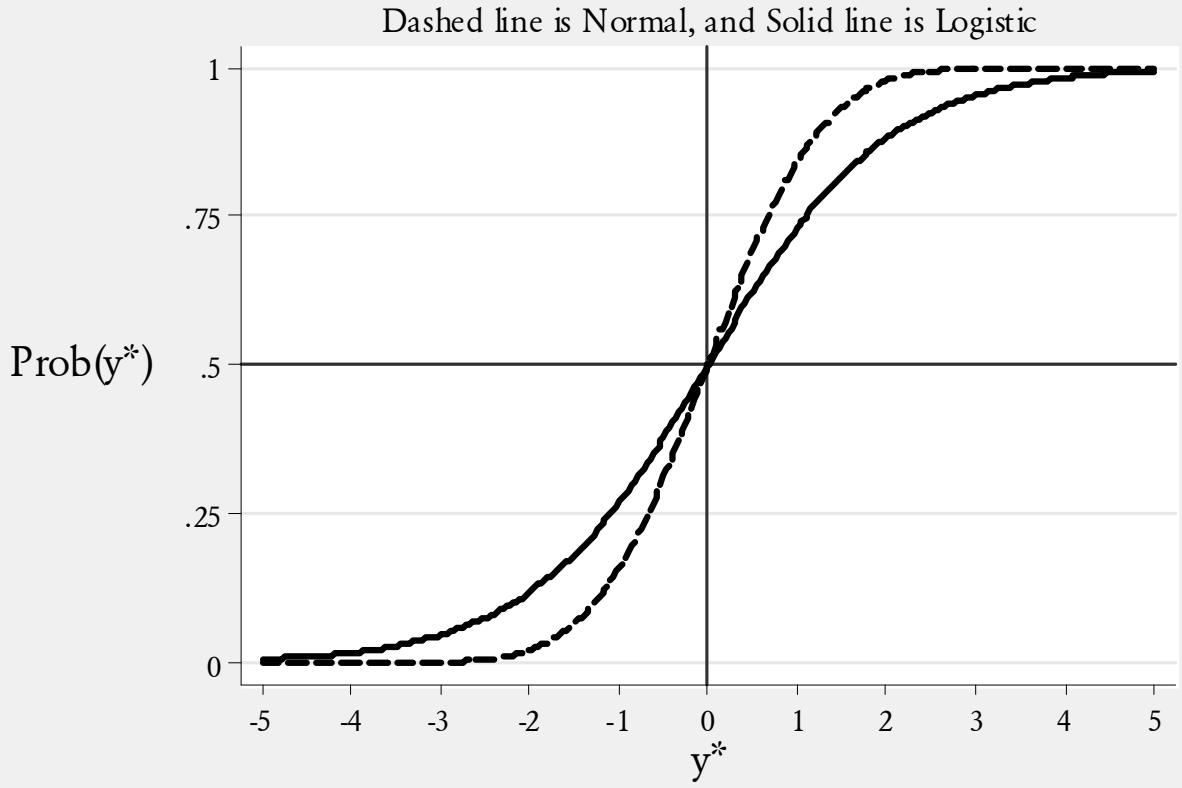
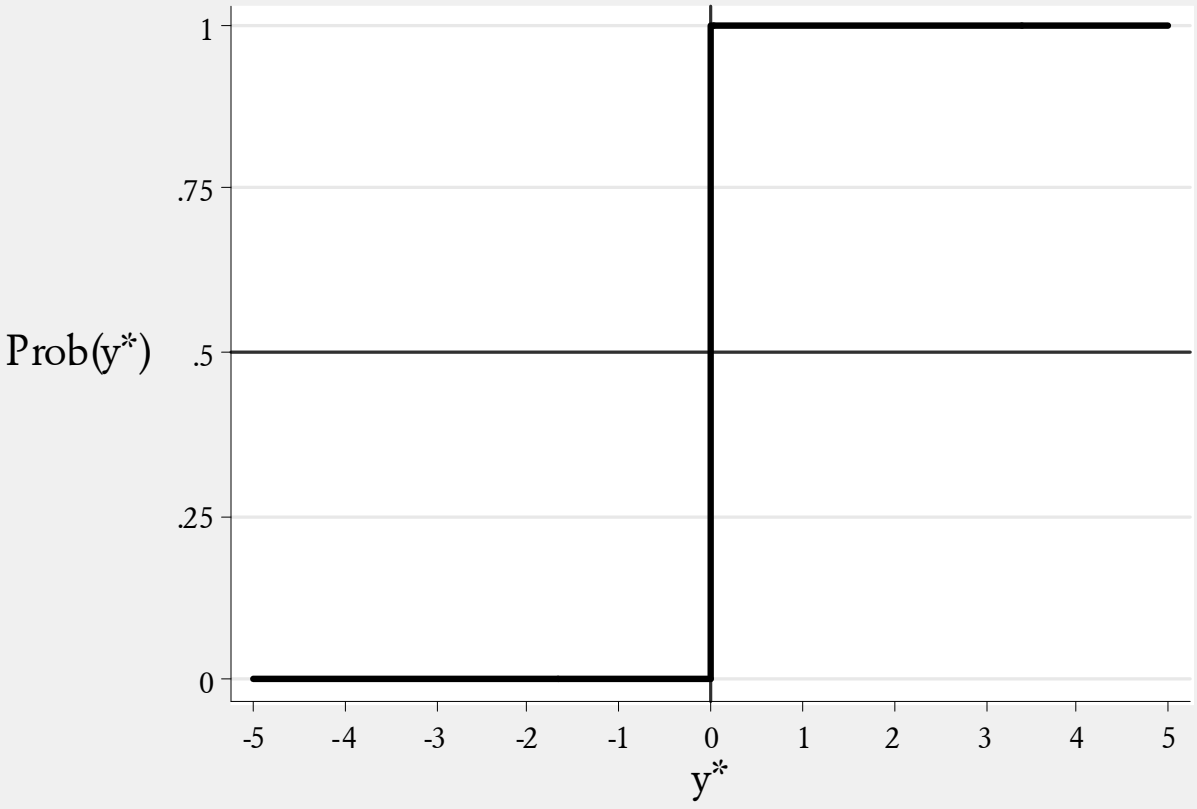


Figure 3: Hardnose Theorist Cumulative Density Function



References

Abdellaoui, Mohammed; Barrios, Carolina, and Wakker, Peter P., "Reconciling Introspective Utility with Revealed Preference: Experimental Arguments Based on Prospect Theory," *Journal of Econometrics*, 138, 2007, 356-378.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Elicitation Using Multiple Price Lists," *Experimental Economics*, 9(4), December 2006a, 383-405.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Valuation Using Multiple Price List Formats," *Applied Economics*, 39(6), April 2007a, 675-682.

Andersen, Steffen; Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet, "Risk Aversion in Game Shows," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 12, 2007b).

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 2008a, 583-619.

Andersen, Steffen, Harrison, Glenn W.; Lau, Morten I.; Rutström, E. Elisabet, "Lost in State Space: Are Preferences Stable?" *International Economic Review*, 49(3), 2008b, 1091-1112.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Preference Heterogeneity in Experiments: Comparing the Lab and Field," *Journal of Economic Behavior & Organization*, 74, 2010, 209-224.

Blackburn, McKinley; Harrison, Glenn W., and Rutström, Elisabet, "Statistical Bias Functions and Informative Hypothetical Surveys," *American Journal of Agricultural Economics*, 76(5), December 1994, 1084-1088.

Camerer, Colin, and Ho, Teck-Hua, "Violations of the Betweenness Axiom and Nonlinearity in Probability," *Journal of Risk & Uncertainty*, 8, 1994, 167-196.

Chambers, Robert G., and Quiggin, John, *Uncertainty, Production, Choice, and Agency: The State-Contingent Approach* (New York, NY: Cambridge University Press, 2000).

Coller, Maribeth; Harrison, Glenn W., and Rutström, E. Elisabet, "Latent Process Heterogeneity in Dicounting Behavior," *Working Paper*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2010.

Conlisk, John, "Three Variants on the Allais Example," *American Economic Review*, 79(3), June 1989, 392-407.

Cox, James C., and Sadiraj, Vjollca, "Small- and Large-Stakes Risk Aversion: Implications of Concavity Calibration for Decision Theory," *Games & Economic Behavior*, 56, 2006, 45-60.

Cox, James C., and Sadiraj, Vjollca, "Risky Decisions in the Large and in the Small: Theory and

Experiment," in J. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Cubitt, Robin P.; Starmer, Chris, and Sugden, Robert, "Dynamic Choice and the Common Ratio Effect: An Experimental Investigation," *Economic Journal*, 108, September 1998, 1362-1380.

Desvousges, William H.; Johnson, F. Reed, and Banzhaf, H. Spencer, *Environmental Policy Analysis With Limited Information: Principles and Applications of the Transfer Method* (New York: Elgar, 1999).

Dufwenberg, Martin, and Harrison, Glenn W., "Peter Bohm: Father of Field Experiments," *Experimental Economics*, 11(3), September 2008, 213-220.

Fiore, Stephen M.; Harrison, Glenn W.; Hughes, Charles E., and Rutström, Elisabet, "Virtual Experiments and Environmental Policy," *Journal of Environmental Economics & Management*, 57(1), January 2009, 65-86.

Grether, David M., and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, 69(4), September 1979, 623-648.

Hansson, Bengt, "Risk Aversion as a Problem of Conjoint Measurement," in P. Gardenfors and N.-E. Sahlin (Eds.) *Decision, Probability, and Utility* (New York: Cambridge University Press, 1988, 136-158).

Harrison, Glenn W., "Experimental Evidence on Alternative Environmental Valuation Methods," *Environmental and Resource Economics*, 34, 2006, 125-162.

Harrison, Glenn W., "The Behavioral Counter-Revolution," *Journal of Economic Behavior & Organization*, 73, 2010, 49-57.

Harrison, Glenn W.; Humphrey, Steven J., and Verschoor, Arjan, "Choice Under Uncertainty: Evidence from Ethiopia, India and Uganda," *Economic Journal*, 120, March 2010, 80-104.

Harrison, Glenn W.; Jensen, Svend-Eric; Pedersen, Lars, and Rutherford, Thomas F., (eds.), *Using Dynamic General Equilibrium Models for Policy Analysis* (Amsterdam: Elsevier; Contributions to Economics Analysis 248, 2000).

Harrison, Glenn W.; Jensen, Jesper; Lau, Morten Igel, and Rutherford, Thomas F., "Policy Reform Without Tears," in A. Fossati and W. Weigard (eds.), *Policy Evaluation With Computable General Equilibrium Models* (New York: Routledge, 2002).

Harrison, Glenn W.; Lau, Morten, and Rutström, Elisabet, "Estimating Risk Attitudes in Denmark: A Field Experiment," *Scandinavian Journal of Economics*, 109(2), June 2007, 341-368.

Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Risk Attitudes, Randomization to Treatment, and Self-Selection Into Experiments," *Journal of Economic Behavior and Organization*, 70(3), June 2009, 498-507.

Harrison, Glenn W.; Lau, Morten I., and Williams, Melonie B., "Estimating Individual Discount Rates for Denmark: A Field Experiment," *American Economic Review*, 92(5), December 2002, 1606-1617.

Harrison, Glenn W., and List, John A., "Field Experiments," *Journal of Economic Literature*, 42(4), December 2004, 1013-1059.

Harrison, Glenn W.; List, John A., and Towe, Charles, "Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion," *Econometrica*, 75(2), March 2007, 433-458.

Harrison, Glenn W.; Rutherford, Thomas F., and Tarr, David G., "Trade Liberalization, Poverty and Efficient Equity," *Journal of Development Economics*, 71, June 2003, 97-128.

Harrison, Glenn W.; Rutherford, Thomas F., Tarr, David G., and Gurgel, Antonio, "Trade Policy and Poverty Reduction in Brazil," *World Bank Economic Review*, 18(3), 2004, 289-317.

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Harrison, Glenn W., and Rutström, E. Elisabet, "Expected Utility *And* Prospect Theory: One Wedding and A Decent Funeral," *Experimental Economics*, 12(2), June 2009, 133-158.

Harrison, Glenn W., and Vinod, H.D., "The Sensitivity Analysis of Applied General Equilibrium Models: Completely Randomized Factorial Sampling Designs," *Review of Economics and Statistics*, 74, May 1992, 357-362.

Hey, John D., and Orme, Chris, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica*, 62(6), November 1994, 1291-1326. **[C]**

Hirshleifer, Jack, and Riley, John G., *The Analytics of Uncertainty and Information* (New York, NY: Cambridge University Press, 1992).

Hoffman, R.R., and Deffenbacher, K.A., "An Analysis of the Relations of Basic and Applied Science," *Ecological Psychology*, 5, 1993, 315-352.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), December 2002, 1644-1655.

Lau, Morten Igel, "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, 2000).

Liang, K-Y., and Zeger, S.L., "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 1986, 13-22.

List, John A., "Scientific Numerology, Preference Anomalies, and Environmental Policymaking," *Environmental & Resource Economics*, 2005, 32, 35-53.

Papke, Leslie E., and Wooldridge, Jeffrey M., "Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 1996, 619-632.

Rabin, Matthew, "Risk Aversion and Expected Utility Theory: A Calibration Theorem," *Econometrica*, 68, 2000, 1281-1292.

Rogers, W. H., "Regression standard errors in clustered samples," *Stata Technical Bulletin*, 13, 1993, 19-23.

Safra, Zvi, and Segal, Uzi, "Calibration Results for Non-Expected Utility Theories," *Econometrica*, 76(5), 2008, 1143-1166.

Saha, Atanu, "Expo-Power Utility: A Flexible Form for Absolute and Relative Risk Aversion," *American Journal of Agricultural Economics*, 75(4), November 1993, 905-913.

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

Stigler, George J., and Becker, Gary S., "De Gustibus Non Est Disputandum," *American Economic Review*, 67(2), March 1977, 76-90.

Wilcox, Nathaniel T., "Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison," in J. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Wilcox, Nathaniel T., "'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice Under Risk," *Journal of Econometrics*, 2010 forthcoming (doi:10.1016/j.jeconom.2009.10.012).

Williams, Rick L., "A Note on Robust Variance Estimation for Cluster-Correlated Data," *Biometrics*, 56, June 2000, 645-646.

Wooldridge, Jeffrey, "Cluster-Sample Methods in Applied Econometrics," *American Economic Review (Papers & Proceedings)*, 93, May 2003, 133-138.