Risk Preferences and Risk Perceptions in Insurance Experiments: Some Methodological Challenges

by

Glenn W. Harrison [†]

January 2024

ABSTRACT

The ability to run experiments, or to see natural data as a quasi-experiment, does not free one from the need for theory when evaluating insurance behavior. Theory can be used to motivate the experimental design, evaluate latent effects from the experiment, or test hypotheses about latent effects or about observable effects that could be confounded by latent effects. The risk, evident in the broader behavioral literature in general, is the attention given to "behavioral story-telling" in lieu of rigorous scholarship. Such story-telling certainly has a role in fueling speculation about possible casual forces at work generating the data we see, but should not be mistaken for the final word. There is also a severe cost in terms of the heroic assumptions needed for identification. Again, such identifying assumptions can have a valuable role, but many general claims rely critically on those assumptions. Controlled laboratory experiments and Bayesian econometric methds should play a complementary role to field experiments and quasi-experiments. One clear lesson from the evaluation of methodological challenges is to use theory more, to explore the ability of "standard economics" to explain behavior. The time has long passed where straw men theories are set up to fail when confronted with behavior. Just as we want to consider flexible parametric functional forms when appropriate, we should be open to conventional economics applied more flexibly.

[†] Maurice R. Greenberg School of Risk Science and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA. Valuable comments from the editors and a reviewer are appreciated. Harrison is also affiliated with the School of Economics, University of Cape Town. E-mail contact: gharrison@gsu.edu.

The decision to purchase insurance is an ideal place to see the economics of risk in action. The demand for insurance pops out of the simplest discussion of the risk premium as the difference between the expected value of a lottery someone faces and their certainty equivalent of that lottery. For a full-indemnity insurance contract with known loss probabilities, no deductible, no coinsurance, and no performance risk, any insurance premium less than the risk premium is a good deal. It is a good deal in the sense that the expected subjective welfare of purchasing the product is positive. That simple theoretical result lends itself to descriptive and normative inferences from experimental data.[1] Even in the setting of experimental or quasi-experimental data, methodological challenges remain with respect to the way in which risk preferences and risk perceptions are incorporated.

This definition of the risk premium is general, and applies whether the decision maker has risk preferences consistent with Expected Utility Theory (EUT), Rank Dependent Utility (RDU) theory, or most alternative models of risk preferences that have any currency within economics.[2] Those alternative theories typically differ on what the implied risk premium is for the same individual, but the logic with respect to the evaluation of demand for insurance is the same.

The RDU model of Quiggin (1982) extends the EUT model by allowing for decision weights on lottery outcomes that can differ from the probabilities of outcomes, and has proven to be one of the most important empirical generalizations of EUT. The specification of the utility function is the same

[1] Harrison (2019)(2024) considers implications for inferences from observational data.

[2] This qualification is just to rule out some bizarre theories of risk preference that have appeared in the economics and psychology literature, and quickly disappeared. Starmer (2000) provides an excellent review of experimental evidence over the years on different models of risk preferences. The controlled laboratory evidence for Cumulative Prospect Theory is actually very weak when reviewed in detail. Moreover, it is common to find empirical studies that claim evidence for probability weighting when literally using RDU specifications for lotteries defined solely over gains, and astonishingly claim that as evidence for Cumulative Prospect Theory: see Harrison and Swarthout (2023). The controlled laboratory evidence is even weaker for Dual Theory, the special case of RDU that assumes a linear utility function. Dual theory plays a key role in identification of "limited consideration" in behavioral insurance when EUT is not assumed, solely because it can be reduced to just one parameter (unlike RDU): for example, see Barseghyan and Molinari (2023).

specification used for EUT. For example, the popular Constant Relative Risk Aversion (CRRA) utility function $U(x) = x^{(1-r)}/(1-r)$ might be used, where x is the lottery prize and $r \neq 1$ is a parameter to be estimated. In this case the parameter r is directly estimating the coefficient of CRRA: $r=0$ corresponds to risk neutrality under EUT, $r<0$ to risk loving under EUT, and $r>0$ to risk aversion under EUT. Let there be J possible outcomes in a lottery. Under EUT the probabilities for each outcome x_j, $p(x_j)$, are those that are induced by the experimenter, so Expected Utility (EU) for lottery i is simply the probability weighted utility of each outcome in that lottery, or $EU_i = \sum_{j=1,J} [\ p(x_j) \times U(x_j)\ ]$. To calculate decision weights under RDU one replaces EU_i with the RDU for lottery i, which is $RDU_i = \sum_{j=1,J} [\ w(p(M_j)) \times U(M_j)\ ] = \sum_{j=1,J} [\ w_j \times U(M_j)\ ]$, where $w_j = \omega(p_j + \ldots + p_J) - \omega(p_{j+1} + \ldots + p_J)$ for $j=1,\ldots, J-1$, and $w_j = \omega(p_j)$ for $j=J$, with the subscript j ranking outcomes from worst to best, and $\omega(\cdot)$ is some probability weighting function. The reason for using rank-dependent weighting is to avoid violations of first-order stochastic dominance if the decision weights had been assumed to be the weighted probabilities, as in $w(p(M_j)) = \omega(p(M_j))$ for all j, rather than the de-cumulative probabilities in RDU. EUT assumes the identity function $\omega(p)=p$.

There are three popular probability weighting functions. The first is the simple "power" probability weighting function proposed by Quiggin (1982), with curvature parameter σ, $\omega(p) = p^{\sigma}$, so $\sigma \neq 1$ is consistent with a deviation from EUT. Convexity of the probability weighting function is said to reflect "pessimism" and, if one assumes for simplicity a linear utility function, generates a risk premium since $\omega(p) < p \ \ \forall p$ and hence the "RDU EV" weighted by $\omega(p)$ instead of p has to be less than the EV weighted by p. The second popular specification is the "inverse-S" probability weighting function used by Tversky and Kahneman (1992), $\omega(p) = p^{\gamma} / (\ p^{\gamma} + (1-p)^{\gamma}\ )^{1/\gamma}$. This function exhibits inverse-S probability weighting (optimism for small p, and pessimism for large p) for $\gamma<1$, and S-shaped probability weighting (pessimism for small p, and optimism for large p) for $\gamma>1$. EUT is the special case $\gamma=1$. The third popular probability weighting function is a general functional form proposed by Prelec (1998) that exhibits considerable flexibility: $\omega(p) = \exp\{-\eta(-\ln p)^{\varphi}\}$, defined for $0<p\leq1$, with $\eta>0$ and $\varphi>0$. When $\varphi=1$ this function collapses to the Power function $\omega(p) = p^{\eta}$, so EUT is the special case $\eta = \varphi = 1$.[3]

The definition of the risk premium also allows for the individual to have subjective beliefs about loss probabilities for insurance, as in Subjective Expected Utility (SEU) where we define π_i as the subjective probability of outcome i in a lottery and simply replace objective probabilities p_i with π_i. The RDU specification can also be directly applied to π_i instead of the objective probability p_i. One then needs to have appropriate priors or data from choice tasks to identify π_i independently of the probability weighting function $\omega(\cdot)$, as demonstrated by Andersen et al. (2014).[4] Hence we should distinguish *conceptually* between someone having subjective probabilities from whether they act "optimistically or pessimistically" towards those (subjective or objective) probabilities. Subjective beliefs about loss probabilities are a challenging confound to many field inferences about insurance, whether or not an experiment was conducted.

The statement that insurance *can* be evaluated by the individual as a risk management tool by the change in the individual's subjective welfare from the decision to purchase the contract is, at one level, not controversial. As a simple matter of theory for standard models of risk preferences in economics, such as EUT, it is uncontroversial, as stressed by Harrison and Ng (2016) and Ericson and Sydnor (2017). But the failure of the assumption that the individual *does* evaluate insurance in this manner, rather than the

[3] Many apply the Prelec (1998; Proposition 1, part (B)) function with constraint $0 < \varphi < 1$, which requires that the probability weighting function exhibit subproportionality (so-called "inverse-S" weighting). Contrary to received wisdom, many individuals exhibit estimated probability weighting functions that violate subproportionality, so it is better to use the more general specification from Prelec (1998; Proposition 1, part (C)), only requiring $\varphi > 0$, and let the evidence determine if the estimated φ lies in the unit interval. This seemingly minor point often makes a major difference empirically. One also often finds applications of the one-parameter Prelec (1988) function, on the fallacious grounds that it is still "flexible" while only using one parameter. The additional flexibility over the Inverse-S probability weighting function is formally valid, but minimal compared to the full two-parameter function. The need to allow for a wider range of probability weighting functions than the Inverse-S is also stressed powerfully by Wilcox (2023).

[4] The additional choice task is a proper scoring rule for eliciting subjective probabilities over binary events, which can be extended to eliciting subjective probability mass functions over non-binary events (Harrison, Martínez-Correa, Swarthout and Ulm (2017)).

statement that the individual *could have* done so, is the basis of many *descriptive* analyses of insurance experiments. And the apparent failure of this assumption as a descriptive matter, along with the possibility that different models of risk preferences might be normatively unattractive, is the basis of many *normative* analyses of insurance experiments. We consider how these apparent failures have been evaluated in field experiments (section 1) and laboratory experiments (section 2), and then review some methodological implications (sections 3 and 4).

There are many types of experiments that can be used to evaluate the economics of insurance. Following Harrison and List (2004), one taxonomy distinguishes laboratory experiments with convenience subjects, artefactual field experiments, and natural field experiments, primarily on the basis of the experimenter's control of the task, the source of subjects, and the awareness the subject has of the experiment. They propose (p. 1013/4) a broad taxonomy to guide understanding of the methodologies in use:

> ... a conventional lab experiment is one that employs a standard subject pool of students, an abstract framing, and an imposed set of rules; an artefactual field experiment is the same as a conventional lab experiment but with a nonstandard subject pool; a framed field experiment is the same as an artefactual field experiment but with field context in either the commodity, task, or information set that the subjects can use; [and] a natural field experiment is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are in an experiment.

We consider experiments in the two extremes of this classification. The presumption throughout, unless otherwise stated, is that the subject faces decisions with real consequences for them, usually but not necessarily financial consequences.[5]

[5] Harrison and List (2004) also consider another category of "thought experiments," which play an important role even with no actual financial consequences since the experiment is not implemented. A wonderful example of a thought experiment becoming a meme, only to be dispelled by someone working out how to actually conduct the experiment, is the claim behind the "calibration critique of EUT" by Hansson (1988) and, much later, Rabin (2000). Cox and Sadiraj (2006) proposed an elegant design to implement a test of this claim, building on the ability to vary "lab wealth" for a given subject, as required from the formal premisses of the claim. Evidence from university undergraduates in the U.S. indicates that the premise is simply false for that population (Harrison, Lau, Ross and Swarthout (2017)), although evidence from representatives of the adult Danish population shows that the premise is

Another taxonomy differentiates experiments by the use of randomization to provide control for unobservable characteristics of subjects, or the use of statistical procedures (e.g., propensity scores, coarsened exact matching, or other matching algorithms) to facilitate "quasi-experimental" evaluations with observational data as if a randomization had occurred.[6] A neglected topic has been the use of such quasi-experimental evaluations from observational data as a source of priors for the efficient conduct of controlled experiments.[7] In economics, a related ethical problem with field experiments arises when scholars propose to "just see what works" without working hard to form priors as to whether the interventions will, in expectation, improve or harm the welfare of the subjects of the experiment.[8]

1. Natural Field Experiments

Many exciting field experiments in insurance (and annuities) have exploited naturally-occurring controls, often with administrative data, and worked hard to augment those data to draw rigorous inferences about the demand for insurance. It is useful to think generally about the methodological approach here, and the assumptions required, and then examine specific instances for natural field experiments with insurance.[9]

valid for the range of lab wealth considered (Andersen, Cox, Harrison, Lau, Rutström and Sadiraj (2018)). In the latter case there are alternative assumptions about the degree of asset integration between field wealth and lottery prizes that allow the reconciliation of small stakes risk aversion with plausible high stakes risk aversion under EUT, and these assumptions appear to apply to the Danish population.

[6] Some *define* an experiment by the use of randomization, but this is far too narrow: see Harrison (2011a)(2013) for discussion.

[7] This is a particularly serious matter with respect to medical procedures and drugs: the long delays in setting up a clinical trial can have potentially dire consequences to patients in the short-term, not least because they form the underwriting basis for many health insurance schemes to cover them. In turn, this problem is exacerbated by the complete disregard of observational data that comes when experiments assume "clinical equipoise" and use none of the insights from observational data in their design, or even as formal priors in sequential trials: see Harrison (2021) for an extended discussion.

[8] This issue is tied up with the recommendations from some that all economics experiments, other than natural field experiments, be pre-registered. The illusory benefits of pre-registration for "good scholarship" aside, few of these pre-registration documents discuss expected welfare effects with any rigor. Normally the presumed beneficial effect on observables is taken as a proxy for doing no harm (in expectation).

[9] Harrison and List (2004; §8) and Harrison (2005) discuss other methodological aspects of natural field experiments.

Some variable or event is said to be a good instrument for unobserved factors if it is orthogonal to those factors. Many of the difficulties of "man-made" random treatments have been discussed in the context of social experiments, which are field experiments commissioned by governments. However, in recent years many economists have turned to "nature-made" random treatments instead, employing an approach to the evaluation of treatments that has come to be called the "natural natural experimental approach" by Rosenzweig and Wolpin (2000).

For example, monozygotic twins are effectively natural clones of each other at birth. Thus one can, in principle, compare outcomes for such twins to see the effect of differences in their history, knowing that one has a control for abilities that were innate at birth. Of course, many uncontrolled and unobserved things occur after birth, and before humans get to make choices that are of any policy interest. So the use of such instruments obviously requires additional assumptions, beyond the *a priori* plausible one that the natural biological event that led to these individuals being twins was independent of the efficacy of their later educational and labor market experiences. Thus the lure of "measurement without theory" is clearly illusory, even in these otherwise attractive settings.

Another concern with the "natural instruments" approach is that it often relies on the assumption that only one of the explanatory variables is correlated with the unobserved factors (Rosenzweig and Wolpin (2000; p.829, fn.4 and p.873)). This means that only one instrument is required, which is fortunate since Nature is a stingy provider of such instruments. Apart from twins, natural events that have been exploited in this literature include birth dates, gender, and even weather events, and these are not likely to grow dramatically over time.

More generally, beyond "nature-made" random treatments, both of these concerns point the way to a complementary use of different methods of experimentation, much as econometricians use *a priori* identifying assumptions as a substitute for data in limited information environments. In turn, that points to the need for formal Bayesian econometric inferences, illustrated later in the discussion of laboratory experiments.

A. Perfectly Informed Risk Types?

A common identifying assumption in many behavioral studies of insurance and annuity choice is that individuals know their own risk type. It is further assumed that it happens to be the risk type that the actuaries at an insurance firm might infer.

Cohen and Einav (2007) examine a rich data-set of choices over menus of deductibles and premium payments for auto insurance that varied across individuals. These menu options constitute necessary controls to view these data as a natural field experiment. The researchers know the premium offered, but do not know the subjective perception of the risk of a claim, or the risk that the claim will be paid in full. To proxy these subjective perceptions they assume that individuals have accurate point estimates of the true distribution, a tenuous assumption even for experienced drivers. Moreover, they must assume EUT, since they have no way to identify non-EUT models of risk preferences, and hence the calibration implications of such preferences. Certain non-EUT models of risk preferences, such as RDU, have been shown to dramatically affect the valuation of insurance when calibrated to estimates from real choices in the field: see Hansen et al. (2016).

This identifying assumption, that individuals know the actuarial loss rates and claim values, turns out to play a critical role in most of the observational literature as well. In a survey Ericson and Sydnor (2017; p.54) correctly note that, "When economists analyze health insurance markets, they typically assume that people are aware of the distribution of their possible medical bills for the year and choose their health plan with that information in mind." In fact, most studies go well beyond assuming awareness of the *distribution*, and are assumed to have statistically degenerate beliefs on some *scalar statistic* derived from that distribution.

Assuming that an individual makes decisions over risky outcomes by reacting optimistically or pessimistically to objective risks is *not* the same as assuming that individuals might have subjective perceptions of risk that deviate from objective risks. Of course, the two might be impossible to tease apart in field settings, but it is easy to do in theory and controlled laboratory experiments that operationalize that theory, as noted earlier. The implications of teasing these apart are apparent when one starts to engage in normative tinkering: one might plausibly adopt a different normative stance towards subjective beliefs being different from the beliefs of some actuaries than the normative stance one takes towards optimism or pessimism with respect to those subjective beliefs.

We can see the difficulty that RDU poses for inference about insurance choice when one allows for subjective probabilities in **Barseghyan et al. (2013)**. They exploit the fact that the decision-makers in their sample had a choice from multiple deductibles, and recognize that this allows them to identify the role of diminishing marginal utility *and* "probability weighting" in the sense of RDU, since these two channels for a risk premium have different implications at different deductible levels. They also explicitly acknowledge that what they call probability weighting might also be simply subjective risk perceptions that differ from the true claims rate, noting that their analysis "does not enable us to say whether households are engaging in probability weighting *per se* or whether their subjective beliefs about risk simply do not correspond to the objective probabilities" (p. 2527). Their striking result is that probability overweighting (or, we add, subjective risk bias) with respect to claims is, along with diminishing marginal utility, a central determinant of the risk preferences of these deductible choices.

A critical assumption tthat they make, common to most of the studies of observational data, is to estimate a *scalar* loss probability for each individual or household in their data. To be sure, these estimates invariably use a rich dataset of demographic characteristics from the data, and presumably available to the actuaries and underwriters of the insurance contract. So they have that level of credibility. But in all cases a point estimate is assumed as if known by the decision-maker, not some subjective probability distribution around that point estimate. To be specific, this assumption is used in Barseghyan et al. (2013; p. 2505), Barseghyan (2021b; p. 1997), Barseghyan et al. (2021a; p. 2028) and in Barseghyan and Molinari (2023; p. 1021). It also plays a key role in the evaluation of health insurance in the Netherlands by Handel et al. (2020; p.11ff.).

B. "Inertia"?

Most insurance contracts have limited contract horizons, usually one year, and are then renewed with potentially different coverage and premia offerings. The behavioral literature often just states that insurees exhibit "inertia," implying that they mindlessly renew contracts even when there appear to be better alternatives available. What might be going on here from the perspective of theory?

Handel (2013) exploits a natural field experiment in which a large firm changed health insurance options from an active choice mode for all existing employees to a passive mode where the previously selected choice was the default choice in later years unless action was taken. This change allowed inferences about the role of "inertia" in insurance plan choice. The behavior of new employees, who needed to make an active choice when previous employees were faced with passive choices as a default, provides intuition for the significance of inertia, assuming comparability of other characteristics between the two employee groups. Some existing employees faced "dominated" choices over time as insurance parameters changed, and their sluggishness in the face of these incentives and the default, dominated option provides indicators of inertia; the use of scare quotes around the term word "dominated" will be explained momentarily.

Risk preferences are assumed to be distributed randomly over the population sampled, and to be consistent with EUT. Individuals know their own risk preferences, but this is unobserved by the analyst. This might cause identification problems if the "nonfinancial attributes," to use the expression of Handel and Kolstad (2015), also varied across all plan choices, but three Preferred Provider Option (PPO) plans had no differences in these attributes: hence their variations in "financial attributes," such as deductible, coinsurance, and out-of-pocket maxima, could be used to identify (atemporal) risk preferences. In keeping with other observational studies, the distribution of claims was simulated using sophisticated models akin to how an actuary would undertake the task, and individuals were assumed to know the risks they faced exactly.

Since the focus is on "inertia" over time, a critical and implicit behavioral assumption is that individuals are *intertemporally risk neutral* with respect to the attributes of the health plan over time.[10] An individual that is *intertemporally* risk averse cares, as a matter of preference, that attributes not vary *over time*. If individuals are assumed to be intertemporally risk neutral then they do not care about variations in attributes over time, as one moves from plan to plan over time, as long as the average attribute remains the same.[11] So giving up their favorite family doctor for a new family doctor does not matter at all, *ceteris paribus* the average attributes of the doctor, and will be accepted willingly for any tiny improvement in premia. For now, assume that the sole attribute considered is the time spent with the doctor, not the identity of the doctor or whether one has a history with the doctor. Then it is being assumed that this

[10] In general the reference to attributes should include what are referred to as "financial attributes" as well as "nonfinancial attributes," but in the context of Handel and Kolstad (2013) the term just refers to the latter. For present purposes the formal theories of multiattribute risk aversion can be viewed as including intertemporal risk aversion as a special case, where one of the attributes is whether the attribute is consumed sooner or later: see Andersen, Harrison, Lau and Rutström (2018). Similarly, one can define multattribute risk aversion even if there is no time-dating of outcomes.

[11] This is separate from the assumption that "consumers are myopic and do not make dynamic decisions whereby current choices would take into account inertia in future periods" (p.2662). That assumption has to do with sophistication with respect to the effect of current consumption on future consumption, akin to "rational addiction" models.

non-financial attribute is the same on average, and the plans can be viewed as dominated on the basis of the financial attributes. The focus here is on an oft-mentioned attribute that, as a matter of fact, was the same across the PPO plans that the individuals being studied could choose from.

But it is clear, as emphasized by Handel and Kolstad (2015; p. 2451) that there is evidence that 50% of subjects did not think the non-financial attributes of the PPO plans were identical, or were not sure of it. All that is needed is that individuals do not *subjectively* believe that these attributes are the same across these PPO plans. This is a false subjective belief: it is not a friction. Given this false belief, the preference for not changing plans could be due to a preference for stability of attributes over time, which is what intertemporal risk aversion is all about. Given this false belief, the plans are not subjectively dominated in terms of the financial attributes. Given this false belief, what is attributed to "inertia" is exactly what a preference for temporal stability implies when one allows for it. And the methodological point is more general, of course, when we consider plan choice over options with objective differences in attributes.

Intertemporal risk preferences are currently modeled in economics and finance in terms of several sharply contrasting structural theories. One imposes intertemporal risk neutrality by assuming an additively separable intertemporal utility function. This assumption is certainly common, but is not fundamental in the same sense that the additivity of the standard Independence Axiom (IA) is for EUT. Similarly, CRRA utility preferences are common, but we would never reject EUT solely on the basis of predictions from a CRRA utility function.[12] This additivity assumption for intertemporal utility also ties atemporal risk preferences and time preferences at the hip, in the sense that they cannot be independent of each other, which seems *a priori* implausible and leads to sharp calibration problems in macroeconomic models.

[12] Well, we *should* not. Gneezy and Potters (1997) did: see Harrison and Rutström (2008; Appendix E).

Various alternative theories allow for some non-additivity in many different way, allowing aversion to stochastic variability over time or a preference for temporally correlated variability. Tolstoy reminded us in the opening line of *Anna Karenina* that "Happy families are all alike; every unhappy family is unhappy in its own way." So it is with additivity and non-additivity. One often finds non-additivity assumed indirectly in terms of "habit formation" models, for example. The specific alternative that we consider to intertemporal risk neutrality, due to Richard (1975), only relaxes the additive separability assumption on the intertemporal utility function.[13]

Define a lottery α as a 50:50 mixture of $\{x_t, Y_{t+\tau}\}$ and $\{X_t, y_{t+\tau}\}$, and another lottery ω at the other extreme as a 50:50 mixture of $\{x_t, y_{t+\tau}\}$ and $\{X_t, Y_{t+\tau}\}$, where $X > x$ and $Y > y$. We can think of $\{x, X, y, Y\}$ as monetary amounts or as non-monetary attributes, and in this context times t and $t+\tau$ are also attributes. Lottery α is a 50:50 mixture of both bad and good outcomes in time t and $t+\tau$; and ω is a 50:50 mixture of only bad outcomes or only good outcomes in the two time periods. These lotteries α and ω are defined over all possible "good" and "bad" outcomes. If the individual is indifferent between α and ω we say that she is neutral with respect to intertemporally correlated payoffs in the two time periods. If the individual prefers α to ω we say that she is averse to intertemporally correlated payoffs: it is better to have a given chance of being lucky in one of the two periods than to have the same chance of being very unlucky or very lucky in both periods. The intertemporally risk averse individual prefers to have non-

[13] The most popular alternative theory is Epstein and Zin (1989) preferences, but they *require* a specific, empirically-rejected, non-EUT structure on atemporal risk preferences, following Dekel (1986) and Chew (1989). These models replace the standard IA of EUT with a Betweenness Axiom (BWA). The difference is easy to explain. The usual IA states that preferences over lotteries A and B are not changed if we consider some lottery consisting of a p chance of A and a (1-p) chance of C and some lottery consisting of a p chance of B and a (1-p) chance of C, for any C and all p. In words, preferences over two lotteries are not affected by adding a common consequence C with the same probability weight. The BWA simply restricts C to be some combination of A or B. The important consequence of this change from the IA to the BWA is that indifference curves within the Marschak-Machina probability simplex are still linear but do not have to be parallel, as in EUT. A significant finding from later experimental work was that linearity *per se* was a descriptive problem, not just *linearity with parallel* indifference curves. One of the most careful reviews and experimental tests, focused directly on the BWA, was from Camerer and Ho (1994), leading Starmer (2000; p. 358) to conclude that a "... general lesson in the data seems to be *don't impose betweenness*."

extreme payoffs *across* periods, just as the atemporally risk averse individual prefers to have non-extreme payoffs *within* periods. One can also view the intertemporally risk averse individual as preferring to avoid correlation-increasing transformations of payoffs in different periods. More formal results, literature review, and experimental evidence that the average Dane is indeed intertemporally risk averse, are provided by Andersen et al. (2018).

In the context of the data evaluated by Handel (2013), intertemporal risk aversion is just a taste for *not* having variability in claims risks over time, where risks refer to all subjective financial and non-financial attributes of the plan, and that is met simply by choosing the same plan year over year. Just as one is willing to pay a risk premium in terms of expected value to reduce atemporal risk aversion, the willingness to put up with lower expected value plans can be seen as a risk premium to reduce intertemporal risk aversion with respect to attributes. This has fundamental implications for the resulting welfare analysis (p.2669-2679). The story here is that "consumers enroll in sub-optimal health plans over time, from their perspective, because of inertia. After initially making informed decisions, consumers don't perfectly adjust their choices over time in response to changes to the market environment (e.g., prices) and their own health statuses" (p.2669). Another story, equally consistent with the observed choices and EUT, is that consumers have a preference for avoiding subjective intertemporal risk in the health plan lotteries they choose. And yet another story has to do with where the false beliefs came from, in this specific context.

C. Risk Preferences Versus Information Frictions?

Handel and Kolstad (2015) seek to tell a story about the role played by "risk preferences" and the role played by "information frictions" in determining the demand for health insurance products. They also seek to tell a story about the welfare implications of the inclusion of "information frictions." I use the expression "seek to tell a story" to be clear that this is academic rhetoric, for the purpose of shifting discussion away from just assuming that "risk preferences" alone explain insurance behavior.[14] Others might not see this type of rhetoric as the right way to model behavior, but that position neglects any appreciation of the paucity of data with which to draw inferences in the field.

Handel and Kolstad (2015) start with a rich administrative data set in which individuals with certain demographic characteristics had to choose between two health insurance plans. One plan, the PPO, provides "comprehensive risk protection" (p. 2451); the other plan, a High Deductible Health Plan (HDHP), provided access to "the same medical providers and treatments as the PPO, lower relative upfront premiums, and larger relative risk exposure." (p. 2451). In addition to the administrative data, for a significant sub-sample of the population they also had a linked survey of beliefs about these plans. The intuition of their results can be seen by one example (p. 2451) referred to earlier: if 50% of individuals incorrectly believed that the PPO provided greater medical access to providers and treatments (20%), or were not sure about that (30%), they were more likely to choose the PPO than individuals that knew that the plans provided the same access. Call these subjective beliefs about some core attributes of the products. Given these subjective beliefs, apply SEU to these choices, and what we see is just a better apple or a less risky apple being selected over a poor apple. The first 20% subjectively perceive a more useful product, and the second 30% subjectively perceive a less risky product.

[14] They reference (p. 2450) Cohen and Einav (2007) and Bundorf, Levin and Mahoney (2012) as conducting welfare analysis of health insurance plans in which they use "observed choices to identify risk preferences." In fact, risk preferences are not identified by Bundorf, Levin and Mahoney (2012). And Cohen and Einav (2007) undertake no welfare analysis. Similarly, Einav et al. (2010a; p. 878) claim that Einav et al. (2010b) and Bundorf et al. (2012) "recover the underlying (privately known) information about risk and preferences." Neither of these are true.

The first formal step in the analysis is just to recover risk preferences from observed choices between the PPO and HDHP. In this case the model assumes EUT, and again *assumes that individuals know the actuarial probabilities* of receiving benefits from each insurance plan. Intuitively, think of the PPO as the safe lottery and the HDHP as the risky lottery.[15] To borrow an expression, the resulting estimates of risk aversion are "just wild and crazy guys," to be laughed at because they are so high (p. 2452). Of course, we know from RDU models of risk preferences that this *might* actually be a combination of (very) pessimistic beliefs about receiving the benefits of the HDHP and a (modestly) concave utility function. The point is that the available data is unable to differentiate these two sources of a risk premium, hence we cannot claim to have identified risk preferences without accepting the maintained assumption of EUT for all individuals, and where EUT assumes remarkably prescient knowledge of the actuarial risks of what are clearly compound subjective lotteries.

The second formal step in the analysis is to correctly recognize (p. 2455ff.) that modern health insurance plans have many attributes that differentiate them. We are not in a world, at least for these product lines, of just trading off lower deductibles for higher premia. In the absence of these "nonfinancial attributes" the utility function has, as an argument, W_k - P_{kj} - s_i where W_k is wealth for household k, P_{kj} is the premium that household k faces for insurance plan j, and s_i is the out-of-pocket payments for some loss event i. Then there is some actuarial probability mass function, let us assume, defined over the s_i, and that depends on the household k and plan j in question. Now consider the effect of "nonfinancial attributes," such as "the network of physicians and hospitals available, the time and hassle costs associated with dealing with claims, and the tax benefits of linked financial accounts." (p. 2455). For short, call this $BLOB_j$ for plan j, recognizing that BLOB has potentially many arguments

[15] The effort to construct these actuarial probabilities (p. 2480) is impressive. It uses *ex post* information to predict the utilization of four types of health expenditure in the coming year, and then *ex post* data on the costs of each of these expenditure types to predict spending distributions. One could use these objective calculations as the basis for eliciting subjective probability distributions with incentive-compatible experiments, which is what we need to estimate an SEU model of insurance choice.

reflecting a vector of perceived attributes.[16] The argument of the utility function then becomes W_k - P_{kj} - s_i + $BLOB_j$. This specification is at the heart of the analysis.

A theoretical problem with this way of handling "nonattribute frictions" is that they are included in an additive manner. This implies that they are known quantities if one knows the household k and plan j, so they are not themselves risky.[17] This further implies that even if they were assumed to be risky, they *cannot* trade off with other "financial risks." The literature on multiattribute risk aversion shows that *additive* utility functions defined over risky attributes exhibits multiattribute risk neutrality, as noted earlier with respect to intertemporal risk neutrality.[18] The general point is that we are talking about "risk preferences" here, albeit in the form of an exciting cocktail of multiattribute risk preferences, but just risk preferences nonetheless.[19]

The modeling upshot is that I am suggesting a different "story" here, and there is no possible way for these data, as rich as they are in comparison to most observational data sets, to tell them apart. But this story has very different implications for how one does descriptive and normative evaluations of observed insurance choices.

[16] Indeed, BLOB could be viewed as a nested utility function defined over these attributes, as proposed in footnote 12 (p. 2456) and in the empirical model. In the empirical model (p. 2475) these attributes are all treated as binary, and included additively.

[17] The only stochastic aspects of these attributes (p. 2456) is that they are *observed* with error by the researcher, reflecting unobserved but deterministic heterogeneity.

[18] It may seem confusing to refer to correlation aversion to attribute variability and intertemporal risk aversion at the same time. The latter just refers to dated risks, where the dating of the risk as "sooner" or "later" is one attribute of the risk and the amount of the payoff on those dates is the other attribute. The usual story of time preferences for non-risky payoffs refers to the trade-off between "smaller and sooner" amounts of money and "larger and later" amounts of money; intertemporal risk aversion just interacts that time-horizon attribute with risk.

[19] Handel and Kolstad (2015; p.2452) include "inertia" in their structural model, and comment that "incorporating inertia into the model matters a lot for risk preference estimates." They refer here to *atemporal* risk preferences. The deeper implications for risk preferences, having to do with *intertemporal* risk preferences, is discussed earlier with reference to Handel (2013), where "inertia" is the main story.

D. Making Dominated Insurance Choices?

Bhargava et al. (2017) study a remarkable data set from a company that offered employees a menu of 48 health insurance plans that differed solely in terms of "financial attributes." In particular, there are blocks of 4 plans that literally differed solely in terms of the deductible and the premium. In one case, Plan A (p. 1329), a $1,204 = $2,134 - $930 increase in the premium was accompanied by a reduction of $650 = $1,000 - $350 in the deductible, and this difference was representative across other plans. Roughly 55% of employees selected a dominated plan, after allowance for after-tax adjustments. Average medical expenditures were $3,567 (p. 1336) and those that chose dominated plans "could have saved an average of $352 with little risk of losing money" (p.1339). In nominal cost terms this is just under a 10% savings compared to expenditures.

Of course, expected savings are not the same as risk-adjusted savings. While it is true that "no beliefs about health care needs or standard preferences for avoiding risk would rationalize the choice of the low-deductible plan" (p.1321), various assumptions could make these welfare losses *de minimis*. An EUT calculation, assuming that individuals again use actual distributions of medical expenditure as their subjective distribution of medical expenditure (p.1342), leads to comparable estimates of the Certainty Equivalent (CE) of the foregone savings. These CE range from $372 down to $167 (p. 1344) depending on the level of risk aversion assumed, as one might expect *a priori*. Of course, an EUT calculation does not take probability weighting into account, even if one continued to assume that subjective expenditure probabilities equaled historical probabilities, and this could have a first-order effect on the implied CE.

Moreover, there is no accounting for aversion to variability of payments over time: a deductible of $1,000 over several years allows more room for variability of out-of-pocket expenditures than a $350 deductible. The same issue arises in the context of the insurance decisions studied by Handel and Kolstad (2015).

A potentially valuable complement to the evaluation of observational data was the use of experiments (in section V) to evaluate alternative explanations in stylized, but "naturalistic" settings. Unfortunately, these were all hypothetical surveys conducted online. These can be useful to set up tests of hypotheses,[20] but suffer from the general problem of hypothetical bias (Harrison (2006)).

2. Laboratory Experiments

A. Normative Evaluation of Insurance Decisions

Harrison and Ross (2023) propose an approach to behavioral welfare economics that is general, and directly applicable to the normative evaluation of insurance purchase decisions. They refer to it as the Quantitative Intentional Stance (QIS):

> Dennett (1971)(1987) provides a rich account of the relationships between beliefs, preferences and propositional attitudes that provides a rigorous foundation for behavioral welfare economics. He argues that the *attribution of preferences and beliefs involves taking an intentional stance toward understanding the behavior of an agent. This stance consists in assuming that the agent's behavior is guided by goals and is sensitive to information about means to the goals, and about the relative probabilities of achieving the goals given available means.* The intentional stance is a product of cultural evolution. It arose and persists because of the importance of coordinated expectations in an intensely social species with massive behavioural heterogeneity due to large brains that support sophisticated learning. Beliefs, preferences, goals, and other propositional attitudes do not have counterparts at the level of brain states. They instead index relationships between target agents, environments, and interpreters trying to explain and anticipate the target agents' behavior (including their communicative behavior). The welfare economist attempting to determine what people regard as subjectively preferable is in the same situation as all people in all social contexts all the time: she seeks accounts of her targets' lattices of propositional attitudes, with particular emphasis on preferences and beliefs about probabilities, that the targets would endorse themselves. She is *not* trying to make inferences about anyone's "latent" states or states that are hidden in brains until someone with a neuroimaging scanner comes along. (p.23/24; footnotes omitted)

Armed with a rigorous theoretical basis for assessing the benefit or harm to an individual from some experimental treatment, how do we make it operational?

[20] In particular, one intriguing hypothesis (p.1353) posits that agents might "value costs associated with plan premiums differently than those paid (perhaps unexpectedly) out-of-pocket." In effect, this relaxes the perfect asset integration assumption that some associate with EUT. Cox and Sadiraj (2006) and Andersen et al. (2018) show how to evaluate partial asset integration specifications using incentivized experiments.

One general recommendation is to use Bayesian methods. The reason that this recommendation is general is that integrating economic theory with experimental data entails the systematic pooling of priors with data, and that is what Bayesian methods are designed to allow. And, critically, one should view the attribution of preferences and beliefs that is central to the QIS as exactly akin to forming priors *about the agent*, and then pooling them with observations *of the agent* to make (normative and descriptive) inferences.

For economists, a canonical illustration of the need to pool priors and data is provided by the evaluation of the expected Consumer Surplus (CS) from observed insurance choices. Even if we limit ourselves to EUT, the gains or losses from someone purchasing an insurance product with known actuarial characteristics depend on their (atemporal) risk preferences. If we have priors about those risk preferences, then we can directly infer if the observed purchase choice was the correct one or not, as illustrated famously by Feldstein (1973). Here the word "correct" means consistent with the inferred EUT risk preferences for the individual making the choice we evaluate normatively. The same point extends immediately to non-EUT models of risk preferences, such as RDU, which can also be used normatively. From a Bayesian perspective, this inference uses estimates of the posterior distributions of individual risk preferences to make an inference over "different data" than were used to estimate the posterior.[21] Hence these are referred to as *posterior predictive distributions*.

In the simplest possible example, considered by **Harrison and Ng (2016)**, subjects made a binary choice to purchase a full indemnity insurance product or not. The actuarial characteristics of the insurance product were controlled over 24 choices by each subject: the loss probability, the premium, the absence of a deductible, and the absence of non-performance risk. In effect, then, these insurance purchase choices are just re-framed choices over risky lotteries. The risky lottery here is to not purchase insurance and run

[21] The *usual* application in Bayesian modeling is to additional out-of-sample instances of the same data used to estimate the posterior. Excellent expositions can be found in Winkler (2003; p. 102 ff.) and Johnson et al. (2022; p. 192ff.). A typical example in the present context would be to predict choices by one of the subjects if she had been offered a new, different battery of choices over risky lotteries.

the risk of the loss probably reducing income from some known endowment, and the (very) safe lottery is to purchase insurance and deduct the known premium from the known endowment.

The same subjects that made these insurance choices also made choices over a battery of risky lotteries,[22] and a Bayesian model can then be used to estimate individual risk preferences for each individual from their risky lottery choices (Gao et al. (2023)). A Bayesian hierarchical model was used in which informative priors for the estimation of *individual* risk preferences were obtained by assuming exhangeability with respect to the risk preferences of *all of the individuals* in the sample. A relatively diffuse (weakly informative) prior was employed to estimate the risk preferences of the pooled representative agent, and the posterior distribution from that estimation was used as the informative prior for estimation of individual risk preferences. One might view the estimates for the pooled representative agent as "nuisance parameters," if they were not so important to the end-result of being able to infer individual risk preferences with informative priors.

Given these posterior estimates of risk preferences for individuals, the task is then to infer the posterior *predictive* distribution of welfare for each insurance choice of each individual. The predictive distribution is just a distribution of unobserved data (the expected insurance choice given the actuarial parameters offered) conditional on observed data (the actual choices in the risk lottery task). All that is involved is marginalizing the likelihood function for the insurance choices with respect to the posterior distribution of EUT model parameters from the risk lottery choices. The upshot is that we predict a

[22] There is a long experimental literature to guide in the selection of experimental elicitation procedures and batteries, as well as appropriate econometric methods for different inferential purposes: see Harrison and Rutström (2008) for a detailed review. Often one sees experimental procedures that offer the illusion of "short cuts" used, despite the problems with those methods being well documented. For example, the procedures of Tanaka et al. (2010) were used by Jaspersen et al. (2022), despite the extensive discussion of problems with those procedures by Harrison and Rutström (2008; p.59ff.).

distribution of welfare for a given choice by a given individual, rather than a *scalar*.[23] We can then report that distribution as a kernel density, or select some measure of central tendency such as the mean or median.

Figure 1 displays several posterior predictive distributions for insurance purchase choices by one subject. For choice #1 the posterior predictive density shows a clear gain in CS, and for choice #4 a clear loss in CS. In each case, of course, there is a distribution, with a standard deviation of $0.76. The predictive posterior distributions for choice #13 and choice #17 illustrate an important case, where we can only say that there has been a CS gain with some probability.

This example allows us to illustrate how one can undertake *adaptive* welfare evaluation during an experiment, following **Gao et al.** (2023; section 3.C).[24] Some of the subjects in this experiment gain from virtually every opportunity to purchase insurance, and sadly some lose with equal persistence over the 24 sequential choices. Armed with posterior predictive estimates of the welfare gain or loss distribution for each subject and each choice, can we adaptively identify *when* to withdraw the insurance product from these persistent losers, and thereby avoid them incurring such large welfare losses? Research by Caria et al. (2023), Hadad et al. (2021) and Kasy and Sautmann (2021) considers this general issue. The challenges are significant, from the effects on inference about confidence intervals, to the implications for optimal sampling intensity, to the weight to be given to multiple treatment arms, and so on.

Assume that the experimenter could have decided to stop offering the insurance product to an individual at the mid-point of their series of 24 choices, so the sole treatment arm was to discontinue the product offering or continue to offer it.[25] The order of insurance products, differentiated by their actuarial

[23] If one was using point estimates from a traditional maximum likelihood approach, or even point estimates from one of the descriptive statistics of a posterior distribution (e.g., mean, median or mode), then the inferred welfare measure would be a scalar.
[24] Harrison et al. (2020b) provide a number of examples of the evaluation of *non-adaptive* treatments, corresponding generally to treatments emphasized in field experiments.
[25] A more sophisticated "targeting" policy might use the information from the first 12 insurance choices to adaptively determine the actuarial parameters that might lead each subject to make better decisions in the remaining 12 choices.

parameters, was randomly assigned to each subject when presented to them. Figure 2 displays the sequence of welfare evaluations possible for subject #1, the same subject evaluated in Figure 1. The two solid lines of Figure 2 show measures of the CS: in one case the average gain or loss from the observed choice in that period, and in the other case the cumulative gain or loss over time. Here the average refers to the posterior predictive distribution for this subject and each choice. Since this is a distribution, we can evaluate the Bayesian probability that *each* choice resulted in a gain or no loss, reflecting a qualitative Do No Harm (DNH) metric enshrined in the *Belmont Report* as applied to behavioral research.[26] This probability is presented in Figure 1, in cumulative form, by the dashed line and references the right-hand vertical axis.

Although there are some gains and losses in average CS along the way, and the posterior predictive probability of a CS gain declines more or less steadily towards 0.5 over time, the DNH probability is always greater than 0.5 for this subject. And there is a steady, cumulative gain in expected CS over time. These outcomes reflect a common pattern in these data, with small CS losses often being more than offset by larger CS gains. Hence one can, and should, view these as a temporal series of "policy lotteries" which are being offered to the subject, if the policy of offering the insurance contract is in place (Harrison (2011b)). In this spirit, we can think of the probabilities underlying the posterior predictive DNH probability as the probabilities of positive or negative CS outcomes, given the risk preferences of the subject. The fact that the Expected Value (EV) of this series of lotteries is positive, even as the probability approaches 0.5, reflects the asymmetry of CS gains and losses in quantitative terms and the policy importance of such quantification. For now, we might think of the *policy maker* as exhibiting risk neutral

[26] See Teele (2014) and Glennerster (2017) for discussion of the *Belmont Report* and some aspects of the ethics of conducting randomized behavioral interventions in economics. Even when randomized clinical trials were not adaptive, or even sequential in terms of stopping rules, it was common to employ termination rules based on extreme, cumulative results (e.g., the "3 standard deviations" rule noted by Peto (1985; p. 33)).

preferences over policy lotteries, but recognizing that the evaluation of the purchase lottery by the subject should properly reflect her risk preferences.

Consider comparable evaluations for four individuals from our sample in Figure 3. Subject #5 is a "clear loser," despite the occasional choice that generates an average welfare gain. It is exactly this type of subject one would expect to be better off if not offered the insurance product after period 12 (or, for that matter and with hindsight, at all). Subject #111 is a more challenging case. By period 12 the qualitative DNH metric is around 0.5, and barely gets far above it for the remaining periods. And yet the EV of the policy lottery is positive, as shown by the steadily increasing cumulative CS. This example sharply demonstrates the "policy lottery" point referred to for subject #1 in Figure 2.

The remaining subjects in Figure 3 illustrate different points: that we should also consider the time and intertemporal risk preferences of the agent when evaluating the policy lottery of not offering the insurance product after period 12. Assume that these periods reflect non-trivial time periods, such as a month, a harvesting season, or even a year. In that case the temporal pattern for subject #67 encourages us to worry about how patient subject #67 is: the cumulative CS is positive by the end of period 24, but if later periods are discounted sufficiently, the subjective present value of being offered the insurance product could be negative due to the early CS losses.[27] Similarly, consider the volatility *over time* of the CS gains and losses faced by subject #14, even if the cumulative CS is positive throughout. In this case a complete evaluation of the policy lottery for this subject should take into account the *intertemporal* risk aversion of the subject, which arises if the subject behaves consistently with a non-additive intertemporal utility function over the 24 periods.

[27] This point has nothing to do with whether the subject exhibits "present bias" in any form. All that is needed is simple impatience, even with Exponential discounting. Berry and Fristedt (1985; chapter 3) stress the importance of time discounting in sequential "bandit" problems in medical settings.

Applying the policy of withdrawing the insurance product after period 12 for those individuals with a cumulative CS that is negative results in an aggregate welfare gain of 108%, implicitly assuming a classical utilitarian social welfare function over all 111 subjects.

One general lesson from this example is that we now have the descriptive and normative tools to be able to make adaptive welfare evaluations about treatments during the course of administering the treatment. How one does that optimally is challenging, but largely because we have not paid it much direct attention in economics. Optimality here entails many tradeoffs, and not just those reflecting the preferences of the instant subject. Our focus here is on the partial equilibrium impact on the *welfare* of each and every individual.[28] This is often confused by economists as trying to evaluate social welfare, a different concept altogether, although ideally concepts that are related to each other in subtle ways. Hence, when we report an average of individual welfare effects descriptively, that is not to impose a utilitarian social welfare function, but just to describe our calculations in a familiar manner. The role of formal general equilibrium welfare evaluations is to account for some of the interactions between agents, and second-best constraints, that affect the evaluation of policy. Just as the numerical models evaluating general equilibrium welfare effects have been extended over the years to include imperfect competition, scale economies, trade barriers that are not *ad valorem* tariffs, and so on, eventually they could be extended to incorporate richer models of behavior in stochastic policy settings. That is not our immediate focus.

The other general lesson from this case study is the difficulty of making decisions during the instant experiment when the inferences from the experiment have some presumed welfare implications for *individuals outside the instant experiment.*[29] If we had truncated these experiments adaptively as suggested,

[28] We stress the welfare impact. Many economists confuse impact evaluation with welfare evaluation, arguing that surely the observable impact being measured must matter for welfare. Even when statistical circumstances are ideal, impact evaluation constitutes at best an intermediate input into the welfare evaluation of interventions. That intermediate input is valuable, but should not be confused with the final product, a proper cost-benefit analysis (Harrison (2014)).

[29] This tradeoff has long been felt keenly in the literature on sequential clinical trials in medicine: see Armitage (1985).

would we have been able to draw reliable statistical inferences about the treatment in a way that would influence future applications of the treatment? The only way to evaluate these issues, particularly with multiple treatment arms, is to undertake them in safe laboratory settings in which subjects literally have nothing to lose, and study the implications of "throwing data away" in accordance with such adaptive rules. Then be Bayesian about deciding how much to learn from that for the potential benefit of society.

B. Methodological Subtleties

The core step in undertaking behavioral welfare economics for insurance decisions is to relax the direct axiom of revealed preferences. It is not, as often thought, just about relaxing conventional assumptions about risk preferences or subjective beliefs, although they play a role in the sequel. This subtle distinction adds to the challenge of undertaking normative evaluation from a behavioral perspective.

Are We Assuming Some True Risk Preference?

No, we are instead assuming some prior belief is being formed about the risk preferences of the agent whose behavior is being evaluated. Thinking of these as priors rather than some "assumed truth" has important implications, quite apart from being consistent with the QIS, and also opens the way to developing ways to better inform the choice of priors for behavioral welfare economics.

The value of viewing these QIS attributions as priors, and employing a Bayesian approach, derives from the methodological need for normative analysis of risky choices to have estimates of risk preferences from choice tasks *other than the choice task one is making welfare evaluations about.*[30] In settings of this kind, it is natural to want to debate and discuss the appropriateness of the risk preferences being used. In fact, the

[30] To be strict, we should say "other than directly, naively inferred from the choice task one is making welfare evaluations about."

need for debate and conversation becomes more urgent when, as here, we infer significant losses in expected CS, and significant foregone efficiency. How do we know that the task we used to infer risk preferences, or even the models of risk preference we used, are the right ones? The obvious answer: we don't. We can only hold prior beliefs about those, and related questions. And when it comes to systematically examining the role of alternative priors on posterior-based inference, one wants to be using Bayesian formalisms.

Saying that we view these as priors is not an invitation to then claim that the welfare evaluation is arbitrary. It is recognizing what economists of a wide range of methodological persuasions have been doing for many decades and just formalizing it. The analogy to the nudge literature is apt. Proponents of nudges correctly stress that when we adopt some choice architecture for decision-makers, and have priors over the effect of that architecture on their behavior, we have simply replaced one existing choice architecture with another. That is, some choice architecture is required, and will be used anyway, so why just assume that historical accident has generated a normatively attractive architecture? Another analogy comes from the the classic *Specification Searches* of Leamer (1978): many of the *ad hoc* methods used by econometricians are clumsy attempts to use priors, so why not recognize that and do it explicitly and elegantly with Bayesian methods?

There are immediate reasons why one would want to use Bayesian estimates of risk preferences for the type of normative exercise illustrated above. One obtains more systematic control of the use of priors over plausible risk preferences, and the ability to make inferences for every individual in a sample.

However, there are also more general reasons for wanting to adopt a Bayesian approach than making explicit the role for priors when making normative evaluations. A related, general reason for a Bayesian approach derives from the *ethical* need to pool data from randomized evaluations and non-randomized evaluations, discussed by Harrison (2021). The motivation for randomized control trials in many areas, such as surgical procedures, derives from non-randomized evidence accumulated in widely varying circumstances, such as the health and co-morbidities of the patient. These data are evidently not inferred from "clean beakers," but they are often *completely discarded* when designing a randomized test of the procedure. This practice reflects the notion of "clinical equipoise," which holds that one should initiate and apply the randomized procedure as if none of the prior non-randomized evidence had existed at all. The counter-argument is just to view those prior data as justifying what is actually observed: someone thinking *a priori* that some new procedure is worth testing. That is not, by construction, a completely diffuse prior at work, so one should formally reflect that fact. The ethical issue takes on urgency when patients are being asked to submit to 50:50 chances of a procedure that these priors suggest is inferior. Of course, such equipoise might be justified by a social objective of arriving at a general conclusion more quickly, for the benefit of all potential patients, despite the expected cost to the instant patient; we would disagree with the implied tradeoff, but we see the logic.

Are We Assuming Stable Risk Preference?

To conduct normative evaluation of insurance decisions in our extended example we needed to make the explicit and necessary assumption that there is a set of risk preferences of an individual that we can identify in a risky lottery task,[31] and that we can apply as priors in an insurance task, so as to infer expected welfare changes from insurance choices. If risk preferences are not stable *over time*, is there a risk of normative evaluation being based on "stale" preferences? If risk preferences elicited in one domain are not stable *across domains*, how do we know that they are appropriate for another domain?

[31] It is true that there are many alternative ways to elicit risk preferences, even if we restrict attention to those that involve incentivized choices. Although content, as a practicing applied economist, with the methods used here (binary choice over a randomly-ordered battery of lotteries), there is no need for anyone to try to define a single "correct" way to elicit risk preferences, and it is not at all clear how to define a sensible metric to use to undertake that race and determine a winner. More strenuously, the existence of different risk elicitation methods is not a reason to pause using one that meets certain attractive criteria, just because there are others under consideration.

Even though these are relevant concerns, we argue that they are second order, simply because there are no other assumptions that one can make *if the objective is normative evaluation*. Now that we have demonstrated the QIS method based on that assumption, however, it is entirely appropriate to engage in debate over the strength or weakness of our prior and potential alternative priors for risk preferences that might be used. This is where the ongoing discussion of these, and related, descriptive characterisations of risk preferences have a legitimate role: helping us navigate among the various priors we might use. In our first attempt at applying the QIS method in the laboratory the risky lottery choice task and the insurance decisions are made contemporaneously, implying that there is no serious issue of temporal stability that arises in this instance. And the financial-outcome frame of the risky lottery choice task is close to the financial-outcome frame of the insurance purchase task, so we also don't anticipate a serious issue of domain-specificity in this instance. But what can we say in general?

Consider the issues raised by any instability of risk preferences over time. Temporal stability of preferences can mean three things, and can be defined at the aggregate level for pooled samples or for individuals.[32] Our concern is with individuals, and this is arguably a more demanding requirement.[33] One interpretation of temporal preference stability is that risk preferences are *unconditionally stable* over time. This means that the risk preference parameter estimates we obtain for a given individual should predict the risk preference parameters she would use in the future when she makes the decision that we are normatively evaluating, no matter what else happens in her life. This is the strongest version of a

These are, after all, just priors. The fact that there are several elicitation methods for these priors does not make the priors *arbitrary*. Rather, it makes them *conditional* on the elicitation methods, and that is all.

[32] The relevant characteristic of stability can also vary with the inferences being made. For some inferences we only care about the ranking of individuals in terms of risk premia, and for some inferences we care about the level of the risk premia for individuals. We assume the latter for our purposes here.

[33] There are very few data collected on any forms of stability at the individual level. Most of the evidence concerns averages or distributions over individuals.

"temporal stability of preferences" assumption, and will presumably be rejected for longer and longer gaps between elicitation of the risk preferences and normative evaluation of the decision.[34]

A second interpretation of temporal preference stability is that risk preferences are *conditionally stable*. This interpretation assumes that risk preferences might be state-dependent and a stable function of states over time, but there could be changes in the relevant states over time. This interpretation implies that the risk preference parameter estimates for a subject might depend on her age, for example, and that particular "state" changes in thankfully predictable ways. Of course, this predictability presumes that we have a decent statistical estimate of the effect of age on risk preferences, but it is plausible that this could be obtained. If the states are readily observable, such as age, conditional stability is perhaps a reasonable prior to have for normative evaluation.

A third interpretation is that risk preferences might be state-dependent and the states are not observable, or that the risk preferences are themselves stochastic. In this instance there are stochastic specifications, which in turn embody hyper-priors, that let us say something about stability (e.g., that the unobserved states are fixed for the individual, or that the stochastic variation in preference realizations follows some fixed, parametric distribution).[35]

Classifying the Type of Risk Preference?

Monroe (2023) takes up a subtlety in the application of the QIS of surprising significance for welfare evaluations. In the application of the QIS by Harrison and Ng (2016) subjects were asked to make

[34] Chuang and Schechter (2015) review the literature and suggest low correlations of risk preferences over time. Harrison et al. (2005) find evidence of unconditional stability over 5 or 6 months for average levels of risk aversion. Andersen, Harrison, Lau and Rutström (2008, §5.1) similarly find evidence of unconditional stability over 17 months for distributions of risk attitudes.

[35] For example, allowing for unobserved heterogeneity Harrison et al. (2020a) find evidence for temporal stability of distributions of risk preferences over 6 to 12 months, but only when correcting formally for sample selection and attrition. And they infer temporal instability when those corrections are not made. No prior study has corrected for selection or attrition when drawing inferences about temporal stability.

a series of binary choices over risky lotteries, to allow inferences about the risk preferences of each individual. Those inferences were, in turn, to be used as priors in the normative evaluation of insurance purchase decisions that the same subjects made in a separate task. A key step in the evaluation was to initially determine if the risk preferences of each individual were better characterized by EUT or by RDU, and then to use the estimates[36] for EUT *or* RDU for that individual when evaluating the insurance choices by that individual. An individual was assumed to be characterized by EUT unless their choices over the risky lotteries exhibited statistically significant evidence of probability weighting, using appropriate tests and various significance levels. In large part, this initial step of "typing" the individual as EUT or RDU was undertaken to make the point that the normative evaluation of insurance choices depends on the *type* of risk preference as well as the *level* of risk aversion.

Monroe (2023) explains that there are two potential problems with this approach. The first problem is that subjects that are determined to be better characterized as EUT decision makers could just be extremely noisy RDU decision makers. And there is no reason to expect that the "noise" in question here affects inferences in some simple additive, linear manner that might wash out when making normative evaluations. The second problem is that declaring somebody to be better characterized as an EUT decision maker, and then using their EUT estimates for normative evaluation, is actually saying that they are a RDU decision maker with *exactly* estimated parameters for their probability weighting function.[37] This is not the same thing as saying that they are "sufficiently noisy" in terms of probability weighting that one cannot reject a null hypothesis that they exhibit no probability weighting at all.

What is remarkable, and demonstrated in numerical simulations by Monroe (2023), is that these seemingly subtle steps in the descriptive characterization of risk preferences can make a significant impact

[36] By "estimates" we mean maximum likelihood point estimates as well as estimates of standard errors and covariances, which were incorporated in the normative evaluation using bootstrap procedures. The Bayesian estimation of risk preferences handles this statistical uncertainty automatically by using posterior *distributions* to characterize all parameters as random variables.

[37] And, of course, that those exact values imply zero probability weighting consistent with EUT.

on normative evaluation. The impacts are most pronounced on the inferred size of the welfare gain or loss, rather than the sign of the welfare effect, but that is not the main methodological point. The important lesson, already adopted in later studies[38] with reference to his arguments, is to just use the RDU model for every subject when undertaking the normative evaluations of welfare. One can still usefully use the classification of risk preference type, by whatever means one wants, to help understand the sources of welfare gains and losses, but one should not use the special case of EUT when an individual is better characterized for *normative* purposes as RDU and EUT is nested in RDU. The clear exception here is if the policy maker or analyst has a well-motivated and explicit reason to maintain EUT as the normative metric for evaluating choices, in which case every subject's risk preferences should be characterized by the EUT model, even if the RDU model does a better job for *descriptive* purposes.

Identifying the Inner Utility Function?

Many who view RDU as a better descriptive model of risk preferences nonetheless view EUT as an appropriate normative model of risk preferences. This raises an important practical issue: if all you have before you as an observer is someone exhibiting RDU behavior, how do you recover the utility function you need to undertake normative evaluations?

One approach is to simply impose EUT on the estimation of risk preferences that are observed, and use the utility function that is then inferred. This approach is used, for purposes of exposition, by Harrison and Ng (2016). One can then argue separately about whether RDU or EUT are appropriate normative metrics to use, and the answer to that argument will generally make a quantitatively and qualitative difference.

[38] For example, Gao et al. (2023) and Harrison et al. (2022).

Bleichrodt et al. (2001) maintain that EUT is the appropriate normative model, and correctly note that if an individual is an RDU (or CPT) decision-maker, then recovering the utility function from observed lottery choices requires allowing for probability weighting and/or sign-dependence. They then implicitly propose using *that* utility function to infer the CE, *but using EUT to evaluate the lotteries*. This is a radically different normative position than the one proposed by Harrison and Ng (2016; p.116).

Some notation will help. Let RDU(x) denote the evaluation of an insurance policy x in Harrison and Ng (2016) using the RDU risk preferences of the individual, including the probability weighting function. They calculate the CE by solving $U^{RDU}(CE) = RDU(x)$ for CE, where U^{RDU} is the estimated utility function from the RDU model of risk preferences for that individual. But Bleichrodt et al. (2001) evaluate the CE by solving $U^{RDU}(CE) = EUT(x)$ where EUT(x) uses the U^{RDU} utility function in an EUT manner, assuming no probability weighting. This is normatively illogical. The logical approach here would be to estimate the "best fitting EUT risk preferences" for the individual from their observed lottery choices, following Harrison and Ng (2016), and then use the resulting utility function U^{EUT} as the basis for evaluating the CE using $U^{EUT}(CE) = EUT(x)$, where EUT(x) uses the same U^{EUT} function used to evaluate the CE.

Modeling Mistakes

Another way to undertake normative evaluations is to develop a structural model of mistakes, and consider the effects on behavior of removing those mistakes. The issue here is whether the modeled behavior is indeed reasonably classified as a mistake or not, and from whose perspective.[39] Behavioral

[39] Measures developed by Alekseev et al. (2024) to implement this idea for models of risk preferences are, methodologically, similar to the Critical Cost Efficiency Index (CCEI) of Afriat (1972), which is used to evaluate the degree of consistency with the Generalized Axiom of Revealed Preference (GARP). The CCEI relative cost measures is defined on the unit interval, and its complement shows what proportion of monetary value an agent should be allowed to waste in order to rationalize her choices by some utility function. While GARP provides qualitative statements, Alekseev et al. (2024) put more structure on the estimation procedure and provide

economists have not been shy to quickly label any "odd behavior" as due to the first heuristic that comes to mind, rather than dig deeper. Harrison (2019; §5.2) provides a critical review of structural models of this kind applied to health insurance and income annuity choices.

An extreme example of this approach is offered by Spinnewijn (2017; p.313ff.), who simply assumes that exogenous frictions exist and that they have labels such as "inaccurate perceptions," "inertia" and "bounded rationality." Handel et al. (2019) claim to be able to estimate these "frictions," and show that they lead observed willingness to pay for health insurance to differ from some inner, "true" valuation of the product. The normative muddle that arises from this approach is evident from Spinnewijn (2017; p.314):

> I assume that *only* the true value is relevant for welfare and policy analysis. Depending on the policy interventions and the frictions considered, some weight could be given to the revealed value as well. For example, in case of inaccurate perceptions, one could argue that when different insurance valuations are caused only by different perceptions of the underlying risk (and not by different perceptions of the actual coverage provided) they should not be considered as frictions at all. [footnote omitted] In case of inertia or bounded rationality, switching or processing costs could be relevant for price policies used to encourage individuals to change contracts, but are arguably irrelevant when mandating an insurance plan. While this caveat should be accounted for in practice, using only true values to evaluate welfare in this stylized framework simply sharpens the contrast with standard *Revealed Preference* analysis.

The omitted footnote is also telling about the confusions here: "See, for example, the subjective expected utility theory in Savage (1954)." SEU is a theory about subjective beliefs, which could just as well be beliefs about the attributes of an insurance contract as about the chances of the insuree having to make a claim about it. In fact, all that SEU does in this case is turn the product itself into a compound lottery, along with the risk that a claim will occur, and the risk that the claim will then be a certain monetary amount coniditional on occurring. The notion that we can partition one of these beliefs as welfare-relevant and not the other beliefs as welfare-relevant is bizarre. And then some "frictions" are welfare-

quantitative evidence of welfare costs from observed choices. This approach can be applied to any model of risk preferences that admits of one or other model of noisy choice.

relevant when someone is choosing a policy but not when the policy is being chosen for them? What type of selective consumer sovereignty is being used here?

Do Structural Models of Risk Preference Predict Insurance Choices?

Some have argued that structural models of risk preferences do not predict insurance choices in laboratory experiments (Jaspersen et al. (2022)), so why should we think that those models provide good priors for normative evaluation? At one level, this criticism simply misses the whole point of doing *normative* evaluations of insurance decisions, and confuses it with a possible *descriptive* goal of trying to predict insurance choices. Or else it confuses the need for *an* attractive prior over risk preferences with the Holy Grail search for *the* "one, true risk preference" for all inferences.

On the other hand, the normative evaluations do already provide clear information on the descriptive ability of risk preferences to predict insurance choices, and with a CS metric that we should care about descriptively as well as normatively. After all, a *positive* CS from the normative exercise already tells us that the risk preference underlying that CS calculation *correctly* predicted the (binary) decision to purchase insurance or not. Similarly, a *negative* CS from the normative exercise already tells us that the risk preference underlying that CS calculation *failed* to predict the (binary) decision to purchase insurance or not.

Of even greater importance descriptively, the normative evaluations tell us when the foregone or gained CS was so small as to be descriptively uninteresting. In many instances, the absolute value of the CS is *de minimis*, accepting of course that one can identify thresholds for CS being unimportant (e.g., one cent, five cents, 5% of some premium, and so on). And the point is just to trace out the effect of different thresholds of interest on inferences, in the spirit of the old "payoff dominance" calculations of rejections of standard theory in experimental economics (Harrison (1989)(1992)(1994)). An implication is that one should never rely on correlations of risk preferences with the insurance *purchase decision* as a metric of predictive accuracy, as in Jaspersen et al. (2022), since many of those decisions might be extremely poorly incentivized.[40] The same is true of linear regressions that look for *additive* associations of multiple risk parameters with the insurance purchase decision, again as in Jaspersen et al. (2022).[41]

3. The Methodological Contrast

The empirical literature in behavioral insurance can be classified into two broad categories from a methodological perspective. One reason for doing this is to point out the advantages and disadvantages of each approach, and to suggest ways that hybrid approaches might mitigate these disadvantages.

The first approach is a "tops down" methodology, illustrated in section 2, that starts with some observed field data that has certain essential features of an experimental design, and asks what identifying restrictions are needed to make certain inferences about behavior. It does not matter if the experimental design was not the intended to aid these inferences: it might just be as simple as the customer being offered a menu of alternative contracts. Indeed, it is often just a menu of insurance contracts with the only

[40] There are many other reasons to avoid such correlations. One is that the exercises in question usually employ point estimates of individual risk preferences, ignoring the imprecision of those estimates. A related concern is that the risk preferences elicited are often interval-censored, such as when they are inferred from a multiple price list, and one should not casually replace these data with mid-points and proceed as if they are not interval-censored (quite apart from how one handles the clopen intervals implied by switches at either extreme of the list). There are standard econometric methods for interval-censored data. Another concern is that it is well-known in elementary statistics that linear correlation can be an unreliable statistic when the underlying relationship is non-linear (Anscombe (1973). Yet another concern is that the correlations refer to one risk preference *parameter* at a time, which of course says nothing in general about the risk *premium* the individual has for the prospect the insurance is designed to manage (e.g., under RDU, one needs to know both the curvature of the utility function and the shape of the probability weighting function to evaluate the risk premium).

[41] There are many other reasons to avoid linear regression. One is that the insurance purchase decision in these experiments is either binary or bounded between 0% and 100%. In neither case is ordinary least squares appropriate, and obviously appropriate methods exist and should be familiar. Another concern is that insurance decisions that allowed the fraction of indemnification to be chosen are very likely to have "spikes" at 0% and 100%, at least under the null hypothesis being tested. In that case (single or double) hurdle models must be used, in conjunction with Beta regression models for the interior decisions. It is incorrect to then "point and click" at a Tobit regression and claim that the results are essentially the same, since that type of regression assumes that the linear latent index can be negative or greater than 100%, which is conceptually impossible.

objective differences being the deductible and the premium of each contract. The advantage of this approach, of course, is that it directly places the researcher and her inferences in the field, in the domain of naturally occurring behavior. The disadvantage is that the identifying restrictions, in terms of risk preferences and subjective beliefs of the customer, often need to be very severe indeed. One of the concerns with this literature is that it often leads to claims that some non-standard behavioral pattern or "friction" has to be at work in order to explain the observed data adequately. The risk here is that the severe identifying assumptions with respect to risk preferences and subjective beliefs might also have explained some or all of those observed data patterns. We highlight the severity of these restrictions in section 2, and link them to alternative hypotheses that could account for the observed data.

The second approach is a "bottoms up" methodology, illustrated in section 3, that starts with some structural theory about how insurance decisions are made, then designs experiments to allow one to identify the "behavioral moving parts" of that structural theory. There is no need for the structural theory to be limited to familiar, standard models of risk preferences or subjective beliefs, but they are often a natural starting place. The strength of this approach is that it directly connects the researcher and her inferences to a structural theory, so that there should be no ambiguity over what the resulting inferences about behavior mean. One limitation of the *application* of this approach is that it is often applied only to convenience sample of university students, even though there have long been "artefactual field experiments" doing exactly the same thing with inconvenient samples that are representative of populations. The use of "auxiliary" artefactual tasks to statistically condition inferences about behavior is the methodological contribution coming from having structural models of behavior, whether that methodological insight is then applied in the laboratory or the field.[42]

[42] Another limitation of the application of this approach is that researchers often use proxy measurements of the risk preferences or subjective beliefs needed, such as hypothetical surveys. In general, these are known from decades of research to be unreliable. This is a limitation of the mis-application of empirical methods, akin to the universal abuse of Ordinary Least Squares estimators in some fields.

Where the "tops down" and "bottoms up" approaches sharply conflict is when we move from descriptive analyses of behavior to normative analyses of behavior (Harrison (2019)). If something is modeled as a "friction" or a "mistake," rather than a preference or a belief, we face different challenges when normatively evaluating behavior. In the former case there is a presumption that removing or overcoming the "friction" or "mistake" will improve welfare for individuals deciding whether to purchase insurance or not. In the later case we need to investigate further if the preference or belief provides a basis for normative inference, as stressed by Harrison and Ross (2018). But it is often the case that the preference or belief are normatively attractive, in the "consumer sovereignty" spirit of welfarism (that welfare judgements should be made on the basis of the preferences and beliefs of the affected individuals). So we quickly end up with sharply different normative implications of the "tops down" and "bottoms up" approaches.

A third approach can be thought of as a hybrid mix of the first two methodologies. In this case we augment the field observations of the "tops down" approach with *priors* about preferences and beliefs from other sources. This is just recognizing "nuisance parameters" from the point of view of statistical identification, and then conditioning on them with non-degenerate priors. For example, one could simply run artefactual field experiments to estimate the preferences and beliefs of samples from the sample population. Or one could conduct auxiliary surveys about beliefs, as in Handel and Kolstad (2015). Or one could use experiments from comparable subjects, with the recognition that these are only comparable subjects, not subjects drawn from the same (target) population. The latter step alerts us to the fact that these are *priors* that the researcher has over the preferences and beliefs of the target population. We can then reasonably discuss what might make better or worse priors for these descriptive or normative inferences, but at least we are focusing on the right idea of a prior rather than magically being able to estimate the "true" preferences and beliefs of the target population, as stressed by by Harrison and Ross (2023).

4. Are We Losing the Risk Management Plot?

A. Just Costs and Prices, Really?

Einav et al. (2010a) apply a "sufficient statistics" approach to measure changes in consumer, producer and social welfare from insurance, based solely on estimates of familiar demand and cost curves.[43] The approach rests on assuming that direct revealed preference applies (p.879), so that the demand curves for insurance products are "sufficient statistics" for willingness to pay for the product. Although consumer welfare and producer welfare are separately identified, there is no basis for making any inferences about the distribution of welfare gains (let alone gains and losses) among the population. Finally, it is limited to evaluating the welfare effects of changes in the *pricing* of *existing* contracts (p.878).

An important by-product of their approach is the claim that their estimates of the expected marginal cost curves of insurers provides a test as to whether information is symmetric or not, and that if there is a rejection of symmetry whether the selection observed is adverse ($MC' > 0$) or advantageous ($MC' < 0$). Unfortunately, these tests rest on the assumed monotonicity of these expected cost curves, which in turn, yet again, depends critically on the assumptions about loss probabilities being the only idiosyncratic characteristic of consumers. Einav and Finkelstein (2011; p.12) recognize this concern:

> More generally, once we allow for preference heterogeneity, the marginal cost curve needs not be monotone. However, for simplicity and clarity we focus our discussion on the polar cases of monotone cost curves.

It is clear that this required assumption leads to a simpler analysis with more-readily available data, but it does not follow that it provides clarity if the required assumption is wrong.

[43] Chetty (2009) provides an excellent exposition of this general methodological approach.

One surprising, related development has been the focus on adverse selection as if it is the *sole* concern for welfare evaluation. **Einav and Finkelstein (2011)(2023)** clearly give this impression. For example, from (2013; p.170): "The distinguishing feature of insurance markets is the cost curve, and, specifically, its link to demand." Really? Perhaps a distinguishing feature of most selection markets, of which insurance is one, but there are many other selection markets other than insurance. And perhaps a distinguishing feature of health insurance and income annuities, but surely not all insurance product lines? And surely not a distinguishing feature of many important insurance contracts, such as index insurance?

The unfortunate upshot of this emphasis is that empirical methods and short-cuts that might be appropriate for measure welfare effects of *adverse* selection in insurance markets are being presented as readily portable to other settings. Again, from (2013; p.173),

> An appealing feature of the [..] framework is that it relies on a standard demand and supply setting, which is a familiar and portable empirical construct. As such, if the researcher has access to the appropriate data and an appealing research design, implementing the [..] framework is reasonably straightforward and can be applied across a range of different insurance markets [...].

Really? We have already noted some issues with the assumptions underlying the approach, and they can and should be debated in the specific contexts (product lines and contracts) in question. But what if one is interested in relaxing the *normative* assumption that demand is driven by direct revealed preference? What if one wants to allow people deciding about purchasing insurance coverage to be making mistakes? Without taking a position (yet) on how one might define, identify and measure those mistakes, are we just *a priori* ruling them out completely in the interests of tractability and portability?[44]

The issues involved in identifying and making welfare evaluations that allow for welfare losses from observed choices are subtle, and were reviewed in section 3 in the context of laboratory experiments

[44] It would be valuable to see extensions to insurance markets of early experimental work on markets with asymmetric information about types, such as Plott, Lynch, Miller and Porter (1984), Miller and Plott (1985), Lynch, Miller, Plott and Porter (1991) and Holt and Sherman (1999). Such controlled experiments make it easier to explore these conceptual issues and their measurement, as well as evaluate the consequences of ignoring certain information when one must compromise with data limitations in the field.

in which they can readily implement it for demonstration purposes and measure welfare losses. Harrison et al. (2020) extend the application of this method to consider a wide range of behavioral interventions proposed in the field to improve take-up and hence, assuming direct revealed preference, welfare, showing that virtually all generate *harm* in terms of welfare losses.

From a normative perspective, a great deal of attention has been devoted to design better insurance *products*. It is apparent from the existing evidence, from the lab and the field, that comparable attention should be devoted to designing higher quality insurance *decisions* from an expected welfare perspective: see Harrison et al (2022b). Of course, what many behavioral economists call better products, worthy of a regulatory nudge here or there, are really better decision scaffolds to facilitate better decisions.[45] We see no real tension here, just the need to have a clear, structured ability to say something about the welfare effect of product innovations *and* the decision process surrounding the product.

B. Forgetting the Risk Management Role of Insurance?

Many evaluations of insurance in the field understand that (formal) insurance is an *ex ante* contract designed to help manage risk. Whether the specific actuarial characteristics of the product do help manage risk or not is a separate matter, and can vary over potential customers. However, it is becoming increasingly common to see evaluations of the provision of subsidies to insurance, or the provision of social insurance, without paying *any* attention to this *ex ante* risk management role. Often this shift in focus is carefully worded, by reference to the evaluation occurring conditional on take-up and the quality of the product. Of course, "conditional on" does not mean "assuming that there are no issues arising from take-up generating welfare losses," a theme stressed in section 3, but it is intended to allow one to get on with the easier job of just looking at correlations of observables.

[45] Taxes, subsidies and purchase mandates, for example, are all decision scaffolds.

The evaluation literature then differentiates between possible *ex post* impacts of insurance and "other" *ex ante* impacts. The *ex post* effects arise because insurance provides a claim payment in the case of a shock, might reduce the likelihood of a household being pushed into a poverty trap, and could prevent them from engaging in harmful coping strategies such as cutting food consumption, taking children out of school and putting them to work, or selling productive assets at moments when the sale value of those assets is low. The "other" *ex ante* effects might arise because insurance could incentivize famers or graziers to increase investment into higher risk production technologies with higher expected returns, leading to increased productivity and income, or make them more likely to take-up financial services such as credit.

These "other" *ex ante* effects and *ex post* effects might be of social value, and probably are. That claim is worth discussion and debate. But the deep risk is that lack of clear evidence for these "other" *ex ante* or *ex post* effects might lead policy-makers to infer that insurance products are not the correct vehicle for managing risk. To be blunt, there might be a wonderful insurance product that delivers the *ex ante* risk management benefits it was designed to do, but delivers none of the "other" *ex ante* benefits and none of the *ex post* benefits. The conceptual confusion here is thinking of insurance as akin to aid, when it is not, and losing sight of a product that nonetheless provides significant expected welfare gains by neglecting to measure them. In practice, it is more likely that it is just easier to get research funding to correlate observables than to do the hard work needed to measure expected welfare.[46]

[46] On the other hand, there is a possible policy rationale for this approach, which might be to mitigate moral hazard when it comes to providing aid. Perhaps a 99% subsidy on an insurance product is just a cunning way to find out which farmers and graziers "really value" protection from risks. Of course, there will be some farmers and graziers that value protection from risks, but do not have the resources or borrowing capacity for the minimal premia. That can be viewed as a second-best constraint in service of the need to avoid wholesale "demands" for aid by those that do not actually "need" it, and hence less aid getting to those that do "need" it. In the nature of this rationale, it must remain unstated lest that statement be used *ex post* to sweep away the veneer of insurance as an *efficient* aid delivery vehicle in a second-best world where we cannot costlessly sort out those that are "most in need" of aid. Clarke and Dercon (2016) explore the tradeoffs of using alternative policy approaches to setting up contingent aid and insurance schemes to manage risks in developing countries facing natural disasters.

5. Conclusions

The ability to run experiments, or to see natural data as a quasi-experiment, does not free one from the need for theory when evaluating insurance behavior. Theory can be used to motivate the experimental design, evaluate latent effects from the experiment, or test hypotheses about latent effects or about observable effects that could be confounded by latent effects. The risk, evident in the broader behavioral literature in general, is the attention given to "behavioral story-telling" in lieu of rigorous scholarship. Such story-telling certainly has a role in fueling speculation about possible casual forces at work generating the data we see, but should not be mistaken for the final word. There is also a severe cost in terms of the heroic assumptions needed for identification. Again, such identifying assumptions can have a valuable role, but all too often we see general claims about the role of "inertia" and "frictions," for example, that rely critically on those assumptions.

One lesson from the evaluation of methodological challenges is to use theory more, to explore the ability of "standard economics" to explain behavior. In a related vein, the time has long passed, one would hope, where straw men theories are set up to fail when confronted with behavior. Just as we want to consider flexible parametric functional forms when appropriate, we should be open to conventional economics applied more flexibly. The conceptual significance of intertemporal and multiattribute risk aversion, even if unfamiliar solely because of the *convention* of assuming additive utility structures, should become standard when considering insurance choices.

The laboratory provides a valuable place to work through conceptual, design and econometric issues, as one aims for field applications. Artefactual field experiments allow much-needed controls for risk preferences, time preferences, and subjective beliefs to be developed to help condition available field data. Lab and field experiments are, as ever, complementary.

Bayesian econometric methods will play an increasingly important role as such auxiliary data are used as priors to condition inferences in the field. In many exciting applications we already see surveys being added, to try to provide insight into latent data-generating processes. We are no longer restricted to use point estimates of nuisance parameters to better condition theoretical inferences for such things as risk preferences and beliefs, and Bayesian methods naturally allow conditioning on prior distributions of parameters.

Figure 1: Posterior Predictive Consumer Surplus Distribution for Each of Four Insurance Purchase Choices by One Subject

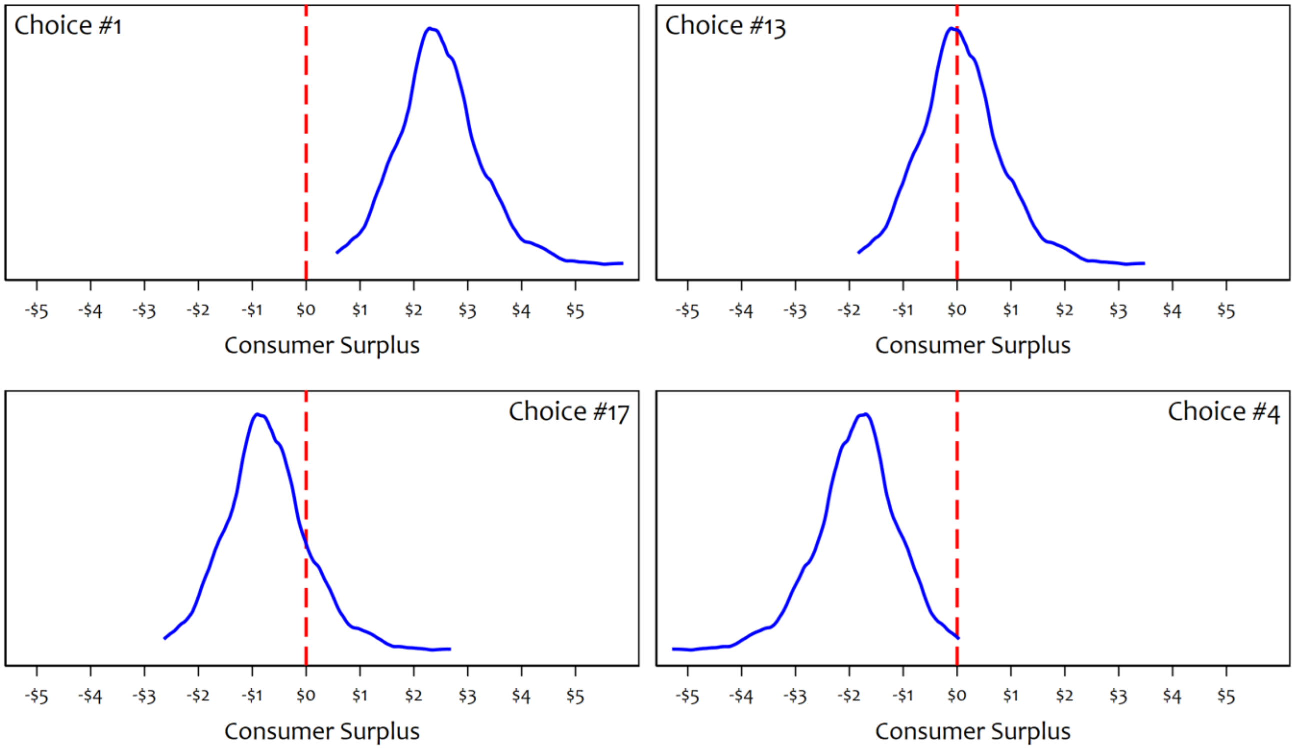


Figure 2: Adaptive Welfare Evaluations for Subject #1

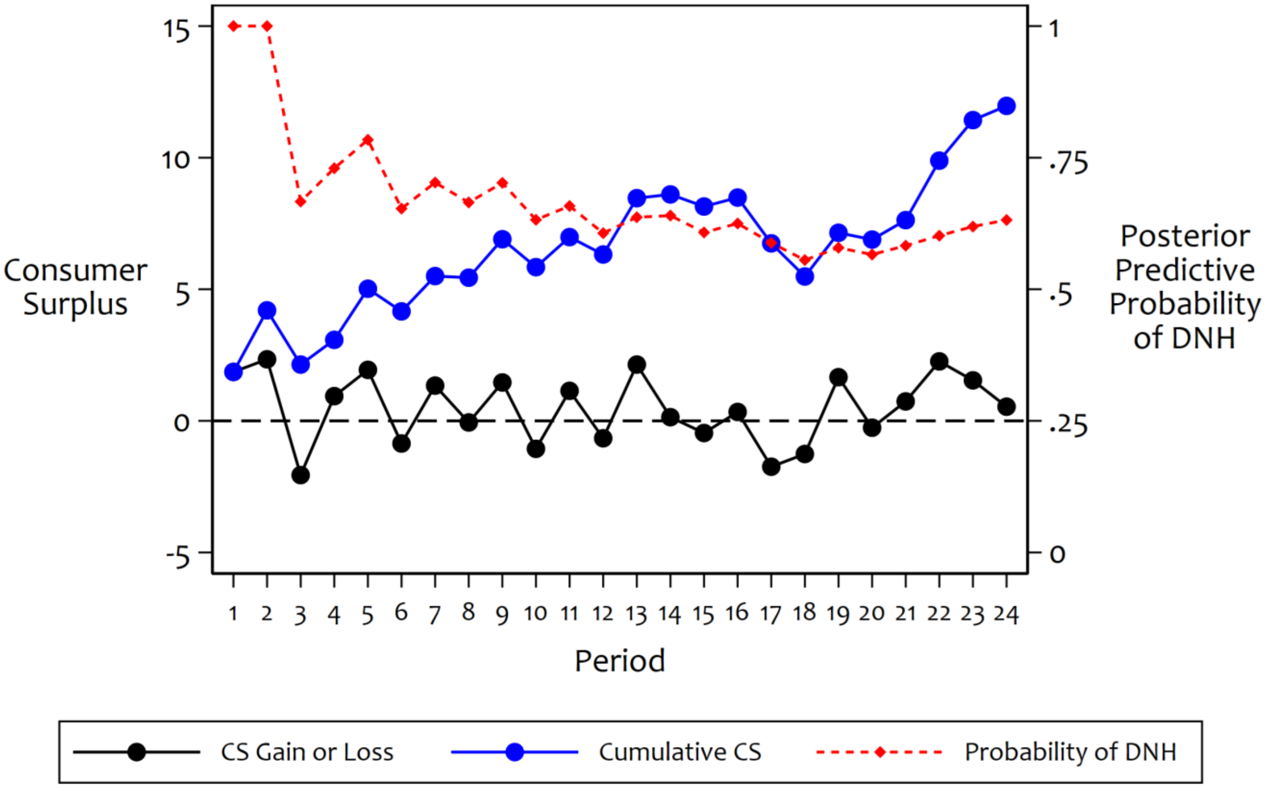
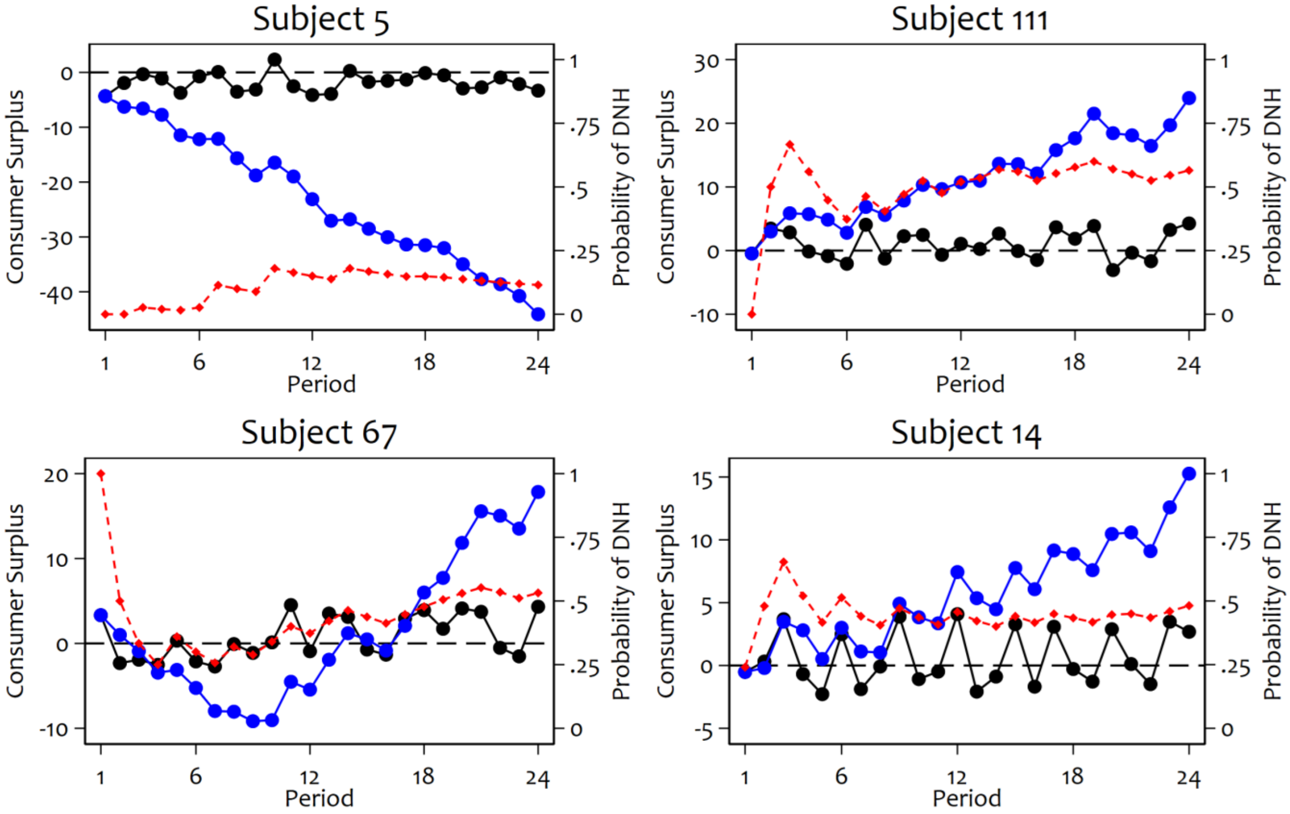


Figure 3: Individual Adaptive Welfare Evaluations for Four Subjects



References

Afriat SN (1972) Efficiency Estimation of Production Function. International Economic Review 13(3):568-598.

Alekseev A, Harrison GW, Lau M, Ross D (2024 forthcoming) Deciphering the Noise: The Welfare Costs of Noisy Behavior. In: Millhouse T, Petersen S, Ross D (eds) Real Patterns in Science and Nature. MIT Press, Cambridge

Andersen S, Cox JC, Harrison GW, Lau MI, Rutström EE, Sadiraj V (2018) Asset Integration and Attitudes to Risk: Theory and Evidence. Review of Economics & Statistics 100(5):816-830

Andersen S, Fountain J, Harrison GW, Rutström EE (2014) Estimating Subjective Probabilities. Journal of Risk & Uncertainty 48:207-229

Andersen S, Harrison GW, Lau MI, Rutström EE (2008) Lost in State Space: Are Preferences Stable? International Economic Review 49(3):1091-1112

Andersen S, Harrison GW, Lau MI, Rutström EE (2018) Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion. International Economic Review 59(2):537–555

Anscombe FJ (1973) Graphs in Statistical Analysis. The American Statistician 27(1):17-21

Armitage P (1985) The Search for Optimality in Clinical Trials. International Statistical Review 53(1):15-24

Barseghyan L, Coughlin M, Molinari F, Teitelbaum JC (2021a) Heterogeneous Choice Sets and Preferences. Econometrica 89:2015–2048

Barseghyan L, Molinari F (2023) Risk Preference Types, Limited Consideration, and Welfare. Journal of Business & Economic Statistics 41(4):1011-1029

Barseghyan L, Molinari F, O'Donoghue T, Teitelbaum JC (2013) The Nature of Risk Preferences: Evidence from Insurance Choices. American Economic Review 103(6):2499-2529

Barseghyan L, Molinari F, Thirkettle M (2021b) Discrete Choice under Risk with Limited Consideration. American Economic Review 111(6):1972–2006

Berry DA, Fristedt B (eds) (1985) Bandit Problems: Sequential Allocation of Experiments. Springer, New York

Bhargava S, Loewenstein G, Sydnor J (2017) Choose to Lose: Health Plan Choices from a Menu with Dominated Options. Quarterly Journal of Economics 132(3):1319-1372

Bleichrodt H, Pinto JL, Wakker PP (2001) Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility. Management Science 47:1498-1514

Bundorf MK, Levin J, Mahoney N (2012) Pricing and Welfare in Health Plan Choice. American Economic Review 102(7):3214-3248

Camerer C, Ho T-H (1994) Violations of the Betweenness Axiom and Nonlinearity in Probability. Journal of Risk & Uncertainty 8:167-196

Caria S, Gordon G, Kasy M, Quinn S, Shami S, Teytelboym A (2023) An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan. Journal of the European Economic Association https://doi.org/10.1093/jeea/jvad067

Chetty R (2009) Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced Form Estimates. Annual Review of Economics 1:451-488

Chew SH (1989) Axiomatic Utility Theories with the Betweenness Property. Annals of Operations Research 9:273-298

Chuang Y, Schechter L (2015) Stability of Experimental and Survey Measures of Risk, Time and Social Preferences: A Review and Some New Results. Journal of Development Economics 117:151–170

Cohen A, Einav L (2007) Estimating Risk Preferences from Deductible Choices. American Economic Review 97(3):745-788

Clarke DJ, Dercon S (2016) Dull Disasters? How Planning Ahead Will Make a Difference. Oxford University Press, New York

Cox JC, Sadiraj V (2006) Small- and Large-Stakes Risk Aversion: Implications of Concavity Calibration for Decision Theory. Games and Economic Behavior 56:45-60

Dekel E (1986) An Axiomatic Characterization of Preferences Under Uncertainty: Weakening the Independence Axiom. Journal of Economic Theory 40:304-318

Dennett D (1971) Intentional Systems. The Journal of Philosophy 68(4):87-106

Dennett D (1987) The Intentional Stance. MIT Press, Cambridge

Einav L, Finkelstein A (2011) Selection in Insurance Markets: Theory and Evidence in Pictures. Journal of Economic Perspectives 25(1):115-138

Einav L, Finkelstein A (2023) Empirical Analyses of Selection and Welfare in Insurance Markets: A Self-Indulgent Survey. The Geneva Risk and Insurance Review 48:167-191

Einav L, Finkelstein A, Cullen MR (2010a) Estimating Welfare in Insurance Markets Using Variation in Prices. Quarterly Journal of Economics 125(3):877-921

Einav L, Finkelstein A, Schrimpf P (2010b) Optimal Mandates and the Welfare Cost of Asymmetric Information: Evidence from the U.K. Annuity Market. Econometrica 78(3):1031-1092

Epstein LG, Zin SE (1989) Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework. Econometrica 57(4):937-969

Ericson KM, Sydnor J (2017) The Questionable Value of Having a Choice of Levels of Health Insurance Coverage. Journal of Economic Perspectives 31(4):51-72

Feldstein MS (1973) The Welfare Loss of Excess Health Insurance. Journal of Political Economy 81(2):251-280

Gao XS, Harrison GW, Tchernis R (2023) Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach. Experimental Economics 26:273-303

Glennerster R (2017) The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency. In: Banerjee A, Duflo E (eds) Handbook of Field Experiments: Volume One. North-Holland, Amsterdam

Gneezy U, Potters J (1997) An Experiment on Risk Taking and Evaluation Periods. Quarterly Journal of Economics 112:631-645

Hadad V, Hirshberg DA, Zhan R, Wager S, Athey S (2021) Confidence Intervals for Policy Evaluation in Adaptive Experiments. Proceedings of the National Academy of Sciences 118(15) e2014602118; DOI: 10.1073/pnas.2014602118

Handel BR (2013) Adverse Selection and Inertia in Health Insurance markets: When Nudging Hurts. American Economic Review 103(7):2643-2682

Handel BR, Kolstad JT (2015) Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare. American Economic Review 105(8):2449-2500

Handel BR, Kolstad JT, Mintem T, Spinnewijn J (2020) The Social Determinants of Choice Quality: Evidence from Health Insurance in the Netherlands. NBER Working Paper 27785, National Bureau of Economic Research DOI: 10.3386/w27785

Handel BR, Kolstad JT, Spinnewijn J (2019) Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets. Review of Economics and Statistics 101(2):326-340

Hansen JV, Jacobsen RH, Lau MI (2016) Willingness to Pay for Insurance in Denmark. Journal of Risk and Insurance 83(1):49-76

Hansson B (1988) Risk Aversion as a Problem of Conjoint Measurement. In: Gardenfors P, Sahlin N-E (eds) Decisions, Probability, and Utility. Cambridge University Press, New York

Harrison GW (1989) Theory and Misbehavior of First-Price Auctions. American Economic Review 79:749-762

Harrison GW (1992) Theory and Misbehavior of First-Price Auctions: Reply. American Economic Review 82:1426-1443

Harrison GW (1994) Expected Utility Theory and The Experimentalists. Empirical Economics 19(2):223-253

Harrison GW (2005) Field Experiments and Control. In: Carpenter J, Harrison GW, List JA (eds) Field Experiments in Economics. Research in Experimental Economics. JAI Press, Greenwich Volume 10:17-50

Harrison GW (2006) Hypothetical Bias Over Uncertain Outcomes. In: List JA (ed) Using Experimental Methods in Environmental and Resource Economics. Elgar, Northampton

Harrison GW (2011a) Randomisation and its Discontents. Journal of African Economies 20:626–652

Harrison GW (2011b) Experimental Methods and the Welfare Evaluation of Policy Lotteries. European Review of Agricultural Economics 38(3):335-360

Harrison GW (2013) Field Experiments and Methodological Intolerance. Journal of Economic Methodology 20(2):103-117

Harrison GW (2014) Impact Evaluation and Welfare Evaluation. European Journal of Development Research 26:39-45

Harrison GW (2019) The Behavioral Welfare Economics of Insurance. Geneva Risk & Insurance Review 44(2):137–175

Harrison GW (2024) The End of Behavioral Insurance. In: Dionne G (ed) Handbook of Insurance, Third edition. Springer, New York

Harrison GW (2021) Experimental Design and Bayesian Interpretation. In: Kincaid H, Ross D (eds) Modern Guide to the Philosophy of Economics. Elgar, Cheltenham

Harrison GW, Johnson E, McInnes M, Rutström EE (2005) Temporal Stability of Estimates of Risk Aversion. Applied Financial Economics Letters 1:31-35

Harrison GW, Lau M, Ross D, Swarthout JT (2017) Small stakes risk aversion in the laboratory: A reconsideration. Economics Letters 160:24-28

Harrison GW, Lau MI, Yoo HI (2020a) Risk Attitudes, Sample Selection and Attrition in a Longitudinal Field Experiment. Review of Economics & Statistics 102(3):552-568

Harrison GW, List JA (2004) Field Experiments. Journal of Economic Literature 42(4):1013-1059

Harrison GW, Martínez-Correa J, Swarthout JT, Ulm E (2017) Scoring Rules for Subjective Probability Distributions. Journal of Economic Behavior & Organization 134:430-448

Harrison GW, Morsink K, Schneider M (2020b) Do No Harm? The Welfare Consequences of Behavioral Interventions. CEAR Working Paper 2020-12, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, Atlanta

Harrison GW, Morsink K, Schneider M (2022) Literacy and the Quality of Index Insurance Decisions. Geneva Risk & Insurance Review 47:66-97

Harrison GW, Ng JM (2016) Evaluating the Expected Welfare Gain from Insurance. Journal of Risk and Insurance 83(1):91-120

Harrison GW, Ross D (2018) Varieties of Paternalism and the Heterogeneity of Utility Structures. Journal of Economic Methodology 25(1):42-67

Harrison GW, Ross D (2023) Behavioral Welfare Economics and the Quantitative Intentional Stance. In: Harrison GW, Ross D (eds) Models of Risk Preferences: Descriptive and Normative Challenges. Research in Experimental Economics. Emerald, Bingley

Harrison GW, Rutström EE (2008) Risk Aversion in the Laboratory. In: Cox JC, Harrison GW (eds) Risk Aversion in Experiments. Research in Experimental Economics, Volume 12. Emerald, Bingley

Harrison GW, Swarthout JT (2023) Cumulative Prospect Theory in the Laboratory: A Reconsideration. In: Harrison GW, Ross D (eds) Models of Risk Preferences: Descriptive and Normative Challenges. Research in Experimental Economics. Emerald, Bingley

Holt CA, Sherman R (1999) Classroom Games: A Market for Lemons. The Journal of Economic Perspectives 13(1):205-214

Jaspersen JG, Ragin MA, Sydnor JR (2022) Predicting Insurance Demand from Risk Attitudes. Journal of Risk and Insurance 90:63–96

Johnson AA, Ott MQ, Dogucu M (2022) Bayes Rules! An Introduction to Applied Bayesian Modeling. CRC Press, Boca Raton

Kasy M, Sautmann A (2021) Adaptive Treatment Assignment in Experiments for Policy Choice. Econometrica 89(1):113-132

Leamer EE (1978) Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York

Miller RM, Plott CR (1985) Product Quality Signaling in Experimental Markets. Econometrica 53(4): 837-872

Monroe BA (2023) The Welfare Consequences of Individual-Level Risk Preference Estimation. In: Harrison GW, Ross D (eds) Models of Risk Preferences: Descriptive and Normative Challenges. Research in Experimental Economics. Emerald, Bingley

Peto R (1985) Discussion of Papers by Bather JA, Armitage P. International Statistical Review 53(1):31-34

Plott CR, Lynch M, Miller RM, Porter R (1984) Product Quality, Consumer Information, and 'Lemons' in Experimental Markets. In Ippolito PM, Scheffman DT (eds) Empirical Approaches to Consumer Protection Economics. Washington, DC, Federal Trade Commission

Prelec D (1998) The Probability Weighting Function. Econometrica 66:497-527

Quiggin J (1982) A theory of anticipated utility. Journal of Economic Behavior & Organization 3(4):323-343

Rabin M (2000) Risk Aversion and Expected Utility Theory: A Calibration Theorem. Econometrica 68:1281-1292

Richard SF (1975) Multivariate Risk Aversion, Utility Independence and Separable Utility Functions. Management Science 22(1):12–21

Rosenzweig MR, Wolpin KI (2000) Natural 'Natural Experiments' in Economics. Journal of Economic Literature 38:827-874

Savage LJ (1954) The Foundations of Statistics, Second edition. Wiley, New York

Spinnewijn J (2017) Heterogeneity, Demand for Insurance, and Adverse Selection. American Economic Journal: Economic Policy 9(1):308-343

Starmer C (2000) Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk. Journal of Economic Literature 38:332-382

Tanaka T, Camerer CF, Nguyen Q (2010) Risk and Time Preferences: Experimental and Household Survey Data from Vietnam. American Economic Review 100(1):557–571

Teele DL (2014) Reflections on the Ethics of Field Experiments. In: Teele D (ed) Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences. Yale University Press, New Haven

Tversky A, Kahneman D (1992) Advances in Prospect Theory: Cumulative Representations of Uncertainty. Journal of Risk & Uncertainty 5:297-323

Wilcox NT (2023) Unusual Estimates of Probability Weighting Functions," In: Harrison GW, Ross D (eds) Models of Risk Preferences: Descriptive and Normative Challenges. Research in Experimental Economics. Emerald, Bingley

Winkler RL (2003) An Introduction to Bayesian Inference and Decision, Second edition. Probabilistic Publishing, Gainesville