Experimental Design and Bayesian Interpretation

by

Glenn W. Harrison [†]

October 2020

H. Kincaid and D. Ross (eds.),
Modern Guide to the Philosophy of Economics (Cheltenham, UK: Elgar, forthcoming 2021).

ABSTRACT

If we take seriously the intent to improve welfare for individuals with experimental interventions, then we must allow that we are also capable of doing harm in terms of welfare. The normative and ethical issues that arise for experimental design and interpretation lead to a derived demand for more economic theory to be used, and rigorous econometric methods. We need to work out how to do behavioral welfare economics in a rigorous and general manner, and we need to consistently apply Bayesian inference methods to avoid the application of dogmatic or *ad hoc* priors. [90 words]

Affiliation: Glenn W. Harrison, C.V. Starr Chair of Risk Management and Insurance, Department of Risk Management and Insurance, and Director, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA. Harrison is also affiliated with the School of Economics, University of Cape Town, South Africa.

ORCID ID: orcid.org/000-0003-1837-8020

Keywords: Experimental design; behavioral welfare economics; Bayesian inference; ethics of experiments.

[†] Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA. Harrison is also affiliated with the School of Economics, University of Cape Town. Valuable comments from Don Ross are appreciated. E-mail contact: gharrison@gsu.edu.

Table of Contents

Experimental methods have grown in importance in economics in the past 60 years. There have been several broad stages in this evolution. In the 1960s there was the use of social experiments to examine major policies in natural settings, in the 1970s there was the use of laboratory experiments to test economic theories in artefactual settings closer to theory, in the 1990s there was the use of field experiments to test economic theories in artefactual and natural settings closer to theory, and in the 2000s there was the use of randomized experimental interventions in developing countries. Along the way, experimental designs evolved to address different question. The appropriate design depends on the question being answered, and the type of inferences to be made.

It is perfectly appropriate for an experimental design *not* to have any randomization at all, such as when one is evaluating whether double-oral auction markets converge to an equilibrium price determined by induced demand and supply curves (e.g., Smith [1962]). Or when one is presenting subjects with risky lottery choices in order to infer risk preferences, and test which theories of risk preference characterize which individuals (e.g., Hey and Orme [1994]). And randomized interventions are not unique to field settings, and have been widely used in laboratory experiments to study the effects of futures markets on the informational efficiency of asset markets (e.g., Forsythe, Palfrey and Plott [1984] and Friedman, Harrison and Salmon [1984]).

It is also perfectly appropriate for an experiment design to be *initially* tethered to some economic theory, such as when selecting parameter values to equalize expected payoffs when evaluating theoretical predictions from single-unit auctions with varying numbers of bidders (e.g., Cox, Roberson and Smith [1982]). Or making sure that the key axioms of models of bargaining behavior are operationalized when testing them (e.g., Roth and Malouf [1979]). And it is appropriate for experimental designs to be motivated by the need to *extend* initial theories as suggested by prior experiments, such as models of sealed-bid behavior (e.g., Cox, Smith and Walker [1984; §V]).

A general concern with experiments spanning this variety of applications in economics is the link between the design of the experiment and the interpretation of results. Increasingly, with specialization and the rise of academic silos, we have separated these. I want to argue for a rejection of that separation, and for the necessity of a Bayesian approach to both. There are two themes to the case for this position.

The first theme is the need for the design and interpretation of experiments not to be divorced from economic theory and the science of econometrics if we are just to do our *descriptive* job well. I have argued for this point elsewhere, and summarize in section 1. The Bayesian approach provides an easily justified contribution to the documentation of empirical regularities and statistical tendencies.

The second theme is the derived demand for a closer connection between the design of experiments and their interpretation for *normative* reasons. If we believe that our experiments might affect the welfare of subjects, or even if subjects believe they might, then we must take this into account in the design phase. By "take into account" it is important to allow for the special case of implicitly ignoring it, since that characterizes many of the practices we observe (e.g., the "equipoise" issue discussed below). And this is referred to as a "derived demand" in part because it rests on developments in behavioral welfare economics, reviewed in section 2, that purport to evaluate the risk of doing harm to subjects from interventions. It is also a "derived demand" because it rests on developments in Bayesian econometrics, reviewed in section 3, that allow us to make statements about the risk of doing harm to subjects at the granular level of an individual choice.

Case studies of medical ethics of clinical trials provide key insights into how these issues tightly connect experimental design and interpretation. Section 4 reviews the debates over the early clinical trials of the Extracorporeal Membrane Oxygenation (ECMO) surgical procedure that provides external support for the heart and lungs with artificial oxygenation of red blood cells. The specific ECMO trials concerned newborn babies in distress, at the point of being close to death without any treatment.

Debates over these trials were brought to the attention of philosophers of science by Worrall [2007].

These insights are connected to practices in the design and interpretation of experiments in economics in Section 5, building on the tools reviewed in Sections 2 and 3.

1. Missing Links to Economic Theory and Econometrics

The claim is that economic experiments need to be closer to economic theory and econometric practice than we now see, even if they are to do their job at describing behavior usefully. To quickly clear the air, if indeed there is debate on these matters, it is useful to summarize points about experimental methodology elaborated on elsewhere:[1]

- I do not concede the causality high ground to randomized evaluations. They have absolutely *nothing* to say about causality statements that entail any cause or effect that is *virtual.*[2] And the big, virtual effect I really care about is welfare, measured as the equivalent (or compensating) variation in income (e.g., Harrison and Ng [2016]). Section 2 is all about how we might measure "welfare" in different ways.

[1] Harrison [2011a][2011b][2013][2014a][2014b][2016][2019][2020], Harrison and List [2004] and Harrison and Ross [2018].

[2] Economists often use the expression "latent states" to mean the same thing. Unfortunately, there are significant complications with the use of the term "latent" when one interacts with philosophers and psychologists, and behavioral welfare economics must interact with them. Ross [2014, §4.2] discusses the complications and why they matter. Harrison and Ross [2018, fn.17, p. 65] summarize the issue as follows: "One way of understanding virtual states is as reaction potentials coupled with environmental affordances in the sense of Gibson [1977], except that the affordances in question will frequently be features of social events rather than (only) features detectable directly by sensory transducers. Because intentional states are propensities inferred from patterns of behavior, they approximately correspond to what some psychologists call 'latent' tendencies. However, psychologists often suppose that latent states have discrete neural realizations that might be discoverable by brain probes or functional neuroimaging. The use of 'virtual' expresses the view among many current philosophers that intentional states generally do not have such realizations because their semantic contents, what is believed or desired or preferred, vary partly with conditions external to the bodies of the agents whose states they are (Burge [1986], McClamrock [1995])." The notion of virtual preferences and beliefs plays a fundamental role in behavioral welfare economics, as explained later.

- The policy relevance of randomized evaluations is never set aside when stressing the need for more "internal validity" in terms of theory and econometric rigor.[3] In fact, it is precisely the *dangerous* policy application of randomized evaluations for policy that gets me agitated about these methodological matters. I reject the view that we should give some research a methodological "get out of jail for free card" just because it claims policy relevance. Nor, conversely, should we *completely* dismiss any research that is policy relevant because it is unable to address *all* methodological issues.
- Lab experiments and field experiments are complementary (Harrison and List [2004]).
- The emergence of "lab-like field experiments" is not recent, and was surveyed in Harrison and List [2004] and called "artefactual field experiments."
- Randomized evaluations have been around for a long time in experiments, as illustrated by the *survey* of then-extant research by Ferber and Hirsch [1978].
- We must use words like "theory" differently if anyone can claim[4] that the randomized evaluation of the effects of the treatment of worms has "contributed significantly to theoretical knowledge" in any form whatsoever, whatever the virtues of that particular empirical study.
- Speculations about possible behavioral mechanisms are not the same thing as identification and measurement of mechanisms, and exhibits the usual selection biases of good (and bad) story-telling (Leamer [1978; ch.10]).
- Pure randomization is, in general, statistically inefficient.
- Using statistical "morning after pills," such as matching methods, to make observational data approximate an experiment design are valuable, but do not obviously generate significant

[3] Bossuroy and Delavallade [2015; p.150] made this claim.

[4] Bossuroy and Delavallade [2015; p. 151].

differences from conventional "regression correction" in realistic applications.

2. Welfare Analysis From the Intentional Stance

There is a large literature on behavioral welfare economics, reviewed critically by Harrison and Ross [2018] and Harrison [2019]. A general concern with this literature is that although it identifies the methodological problem well, no contributions provide "clear guidance" so far to practical, rigorous welfare evaluation[5] with respect to risk preferences as far as we can determine. That is what the approach advocated by Harrison and Ng [2016] and Harrison and Ross [2018] seeks to do. They use the best descriptive model of risk preferences to make normative evaluations of the insurance product choices or alternative investment portfolio choices by their subjects.[6] The choice of this approach is evidently of direct relevance with respect to the extent of paternalism involved in normative assessment, and can be justified on deeper philosophical grounds.

Harrison and Ross [2018] included a case study from a consulting project undertaken for an investment bank. Based on evidence that RDU choosers suffered significantly more welfare losses than EUT choosers, they recommended additional cognitive preparation for RDU choosers before they selected investment products, but did not recommend trying to teach them the concept of probability weighting so they could then apply this characterization to themselves. This is only partly motivated by the questionable practicality of the pedagogical task that would be required. It also reflects wariness about telling subjects a story about themselves they would surely interpret as telling them that they possess a kind of internal psychological "defect" when such a story would outrun the available data and is, in any case, doubtful according to sophisticated philosophy of mind.

[5] Sugden [2018] puts the argument for a "coherent" approach to normative economics, challenging the belief that we can do much more than that.

[6] The exposition in this sub-section is adapted from Harrison and Ross [2018; §5].

It is unlikely that most people choosing insurance contracts or investment funds attempt to compute internally represented optima, either from EUT or RDU bases, and then make computational errors that could be pointed out to them. This echoes a point made by Infante, Lecouteux and Sugden [2016] when they complain that behavioral welfare economists typically follow Hausman [2011] in "purifying" empirically observed preferences. Infante et al. [2016] argue that purification reflects an implicit philosophy according to which an inner Savage-rational agent is trapped within a psychological, irrational shell from which best policy should try to rescue her. They provide no general philosophical framework within which they motivate their skepticism about "inner rational agents." However, such a framework is available.

Dennett [1987] provides a rich account of the relationships between beliefs, preferences and other "propositional attitudes" that provides a rigorous philosophical 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. Goals, like preferences and beliefs, are *not internal states of agents*, but are rather relationships between agents, environments, and those of us that are attributing these relationships in order to rationalize voluntary behavior. Behavioral welfare economics involves precisely such rationalizations. Hence there is a crisp rejection at the outset of the internalist conception of economic agents presented in naïve behavioral rhetoric, such as the "humans" *versus* "econs" contrast.

The behavioral welfare economist, by this account, has to try to interpret and predict the agent's actions by means of controlled speculation about that agent's context and information-processing capacities. Agents themselves are trained, during socialization while growing up, to adopt the intentional stance toward themselves. For the sake of coordination in action and communication, agents' self-ascriptions are made so as to at least approximate alignment with the ascriptions of others. These ascriptions and self-ascriptions are not guesses about "true" beliefs and preferences hidden from direct view in people's heads. Rather, beliefs and preferences are constructed rationalizations of agents' behavioral and cognitive ecologies.[7]

Beliefs and preferences are virtual states of whole intentional systems rather than particular physical states of brains; but being virtual is a way of being real, not a way of being fictitious. If a claim about intentional states is the sort of claim that can have a truth value, then it had better be possible to specify possible evidence that would undermine it. The holistic nature of intentional stance description of agent behavior allows for error, but also complicates it: as stressed by Hey [2005], the "behavioral error" stories that we append to our structural models are part of the economics.[8]

Ross [2014] argues that this marks a main basis for the distinction between economics and psychology. Psychologists are professionally interested directly in how individuals process information, including information that influences decisions. Economists, by contrast, are concerned with this only derivatively. If a system of incentives will lead various people, through a heterogeneous set of psychological processes, to all make the same choice then the people form, at least for an analysis restricted to that choice, an equivalence class of economic agents. But it is a strictly empirical matter when this psychological heterogeneity will and won't matter economically. Economists, like all

[7] Critics have sometimes misinterpreted this view as instrumentalism, a doctrine according to which beliefs and preferences are mere useful fictions, unconstrained by "facts of the matter." Dennett [1987] has consistently maintained, however, that there are facts about agents' goals and access to information, and hence also facts about their propositional attitudes, that should constrain these rationalizations. And these facts are testable by out-of-sample predictions.

[8] To add complication, they interact directly with the stochastic specifications that attend to sampling errors in the econometrics, and hence inferences about preferences: see Wilcox [2008] for a masterful review in the case of risk preferences.

scientists, seek generalizations that support out-of-sample predictions. Different data-generating processes tend to produce, sooner or later, different data, including different economic data. Economics is thus crucially informed by psychology in general, while not collapsing into the psychology of valuation as some behavioral economists have urged (e.g., Camerer, Loewenstein and Prelec [2005]).

Applying this philosophy of mind and agency to the applications to insurance in Harrison and Ng [2016], we assume the intentional stance to make sense of our experimental subjects' overall behavioral patterns, and use the lottery choice experiment as a relatively direct source of constraint on the virtual preference structures we assign when we perform welfare assessment of their insurance contract choices. The more precisely we specify the contents of propositional attitudes, especially in quantitative terms, the less weight in identification will rest on "inboard" elements of data generating processes relative to external aspects of the agents' overall behavioral ecologies (i.e., cognitive scaffolds). Our technical tools allow us to identify virtual intentions that most subjects are not able to identify when they take the intentional stance to themselves, and that they could not deliberately use to evaluate their own decisions.[9] On the other hand, certain experimental treatments[10] might provide evidence that attention to certain informational patterns induces a significant number of subjects to act as if they were stochastically closer to expected utility optimizers. These patterns therefore enter into a fully informed analyst's specification of the subjects' beliefs and preferences.[11]

Application of the intentional stance to economics seems to demand use of Bayesian

[9] Hence, again, the irrelevance of the derisive comments of some behavioral economists towards their straw man account of the agent being modeled, on the grounds that nobody actually makes decisions the way our intentional stance posits.

[10] For example, the informational treatment of Harrison and Ross [2018] with respect to investment decisions.

[11] In this philosophical framework, it makes sense to say that we *boost* the subjects' informational access in a way that *nudges* their (sub-deliberative) cognition.

reasoning.[12] The reason is that it relies on the widespread attribution of preferences and beliefs to agents, along with controlled speculation about that their history, immediate context and information-processing capacities. These attributions must be able to be revised continuously as new data on the behavior of the agent, or the context that the agent is in, comes along. Only Bayesian methods allow one to undertake such inferences consistently over time.

Coupling the intentional stance with Bayesian reasoning also allows some insight into possible explanations for situations in which the intentional stance has been claimed to fail. Some biologists have denounced it as "anthropocentrism" in its application to non-human animals. If this criticism is valid in general then its logic must undermine many applications to humans.[13] But all that actually follows from this concern is that the intentional stance should not be carelessly applied with dogmatic priors. Some ethologists, such as Seyfarth and Cheney [2002], have usefully applied the intentional stance after amassing data on particular species and weighing the quality of different data carefully. We will see many instances in sections 3 and 4 of the need to weigh evidence from different sources, and to make judgments about the "exchangeability" or comparability of those sources, when we work through implications of the intentional stance for experimental design and inference. Only Bayesian methods allow us to pool different sources of information in a systematic way.

[12] The application of Bayesian reasoning here is to the application of the intentional stance by researchers. Discussion of the role of Bayesian hierarchical models as themselves being a "competence model of the brain" can be found in Dennett [2016; ch. 8], from whom the quote is extracted, Hohwy [2013] and Clark [2015].

[13] For additional discussion see Dennett [1983] and the ensuing debate over that "target article," and Bogdan [1997].

3. Bayesian Econometrics

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

A. Examples

As an initial example, Wilcox [2015] provides a *tour de force* of forensic methodology, revisiting the details of the famous Millikan oil-drop experiments. He translates each of the steps, and mis-steps, that Millikan went through in his experimental design and statistical modeling into language that we can contrast with the "credibility revolution" in economics.[14] Unlike many of the commentators on statistical issues surrounding randomized evaluations, we see a clear focus on the interplay between theory, experimental detail, and statistical assumptions. The evaluation of Millikan is important because many of the defenses of the methods and practices of randomized evaluations are *ad hominem*.[15] Consider also the final comments of Johnson [2008; p.156], who referred to the Millikan experiments as one of *The Ten Most Beautiful Experiments* :

> More interesting than the unfounded allegations is the question of how you keep from confusing your instincts with your suppositions, unconsciously nudging the apparatus,

[14] For comparable translations, also see Leamer [2010].

[15] For instance, am I the only one to bristle at Bossuroy and Delavallade [2015] when told that points about statistical methods are not worth discussing because "Arguments regarding the RCTs' statistical properties have already been thoroughly discussed by authoritative figures in the field"? In fact, it is the separation of discussion about "statistical properties" of an estimator and discussion of welfare-theoretic interpretation of that estimator that is one of my major themes. There is more at stake here methodologically than my Australian rejection of Colonial Rule authority. The gains from academic specialization are only realized through gains from trade between the specializations.

> like a Ouija board, to come up with the hoped-for reply. It's something every experimenter must struggle with. The most temperamental piece of laboratory equipment will always be the human brain.

We are indebted to Wilcox [2015] for this careful exegesis, reminding me to once again try to take my Bayesian roots more seriously.

An immediate illustration of the need to pool priors and data is provided by the evaluation of the expected consumer surplus (CS) from observed insurance decisions. Even if we limit ourselves to EUT, the gains or losses from someone purchasing an insurance product with known actuarial characteristics depends on their risk preferences. If we have priors about those risk preferences, then we can directly infer if the observed purchase decision was the correct one or not. The same point extends immediately to non-EUT models of risk preferences, such as Rank-Dependent Utility (RDU). 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.[16] Hence these are referred to as *posterior predictive distributions*.

In the simplest case, 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: the loss probability, the premium, the absence of a deductible, and the absence of non-performance risk. In effect, then, these insurance purchase decisions are just re-framed choices over risky lotteries. The risky lottery here is to not purchase insurance and run the risk of the loss probability 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

[16] The usual application in Bayesian modeling is to additional out-of-sample instances of the same data used to estimate the posterior. A typical example would be to predict choices by one of our subjects if she had been offered a new, different battery of choices over risky lotteries.

lotteries, and a Bayesian model can be used to estimate individual risk preferences for each individual.[17] So the task is 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 model parameters from the risk lottery choices. The upshot is that we predict a *distribution* of welfare for a given choice by a given individual, rather than a *scalar*.[18] 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 decisions by one subject. For decision #1 the posterior predictive density shows a clear gain in consumer surplus, and for decision #4 a clear loss in consumer surplus. In each case, of course, there is a distribution, with a standard deviation of $0.76. The prediction posterior distributions for decision #13 and decision #17 illustrate an important case, where we can only say that there has been a consumer surplus gain with some probability. We return to this application for additional inferences in Section 5.

B. The General Case for Bayesian Methods

In summary, there are immediate reasons why one would want to use Bayesian estimates of risk

[17] Details are provided in Gao, Harrison and Tchernis [2020]. 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 other individuals in the sample. A diffuse prior was employed to estimate the risk preferences of the representative agent, and the posterior distribution from that estimation used as the informative prior for estimation of individual risk preferences.

[18] 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.

preferences for the type of normative exercise illustrated here: 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, to make explicit the role for priors when making normative evaluations.

A general example of this need can be seen in randomized clinical trials, where we often have "clean beaker science" being applied in a "dirty beaker world." Consider drug approvals, another one of those gold standard *ad hominen* references. Phase III trials often control carefully for co-morbidities and confounds: an atypical antipsychotic for manic episodes for bipolar type II might be tested on adults with no financial difficulties, no erectile dysfunctions, and no other other chronic conditions. During the trial adherence to the prescription protocol is enforced. And the evaluation of side-effects in phase III trials lasts how many weeks, months or yeasrs? But then the drug is approved for use in the naturally-occurring world of patients with these confounds, who may not follow the prescription protocol, for drugs that are prescribed for years, and we must rely on a controversially secretive "adverse effects" post-approval reporting process to check for problems (understandably) not detected in the clean beaker. And medical doctors have lobbied for the right to prescribe off-label, and of course do so. I do not have a better system to suggest, but don't tell me that we should promote an illusion of risk regulation and safety here on the back of the clean beaker science. And please do not tell me that this is an authoritative gold standard to which you want me to hold my research. But do allow me to encourage more careful and conditional Bayesian evaluations of the risk of doing harm when "exhangeability" is not obvious between the priors from the lab and regulatory inferences for the field.

Another general reason for a Bayesian approach derives from the ethical need to pool data from randomized evaluations and non-randomized evaluations, as discussed in Section 4. Another general reason for a Bayesian approach derives from the methodological need for normative analysis to have estimates of risk preferences from choice tasks *other than the choice task one is making welfare evaluations about*. 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 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 is: 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.

Here is an example to illustrate this general point. Imagine one was designing a field experiment, say in rural Ethiopia, in which various interventions for a health insurance product were to be used to improve welfare. Assume a health insurance product focused on acute conditions, with significant mortality risk. The only priors on risk preferences you have come from university students in the United States. Should you go ahead and design interventions that, conditional on those risk preferences, lead to welfare losses for the same students, of the kind we have demonstrated? We suggest that, ethically speaking, you should not.

Now imagine you have been able to conduct comparable artefactual field experiments over *money* in Ethiopia that allow you to infer risk preferences, and assume that these experiments match the standard criteria we have for taking any experimental data seriously (e.g., financial incentives and incentive compatibility). These are obviously better priors for the eventual inference, and should be used. You completely discard the priors from students in the United States, or give them relatively lower weight in your hierarchical priors.

Then imagine that you have been able to conduct artefactual field experiments over *certain* health outcomes in Ethiopia that allow you to infer risk preferences. Assume that these health outcomes refer to morbidity risks, not mortality risks, but to real outcomes nonetheless. As any experimental economist knows, it is not easy to come up with morbidity outcomes that can be credibly and ethically delivered within the budgets we normally find ourselves constrained by. Clearly the domain of risk preferences here is *closer* than the risk preferences defined over money, but would you now attach zero or negligible weight to the risk preferences over money by similar Ethiopians? Probably not. So how do you pool these priors to arrive at inferences? The answer is to be Bayesian.

4. Insights from Some Debates Over Medical Ethics in Clinical Trials

The availability of metrics for the evaluation of welfare effects of policy interventions in economics implies that more attention should be paid to the ethical risk of doing harm to subjects. This assumes that these metrics admit of harm at all, and that is not typically the case. But assuming that such metrics are available, a series of questions from the older medical literature on clinical trials arises. Are we free to evaluate any policy intervention, even in the absence of "hard data" to guide us? Is it appropriate to randomize to treatment arms equally? Should there be any adaptation in the weights assigned different treatment arms during the trial? What implications flow for the consenting process? Should we consider adaptive termination of a harmful treatment for a specific subject if there are repeated exposure to that treatment?

For ethical reasons, it is almost always necessary to pool data from randomized evaluations and non-randomized evaluations.[19] The ethical need arises prior to conducting any experiment that involves

[19] Quite apart from the ethical need, it is common sense to use all available information in descriptive analyses. Encountering economists that only believe what a field experiment tells them, and ideally what a natural field experiment tells them when serendipity bestows a discontinuity amendable to regression, it is useful to remind them of the importance of "lab experiments" to Darwin. The first chapter of *Origins* was mainly about domestication and artificial selection, from which he formed priors for the field and the data he had on natural selection. The parallels between methodical, unconscious and natural selection were explored even more formally in Darwin [1868].

randomized treatments, and then during and after the trial.[20]

A. Experimental Design Prior to the Experiment

The first ethical design issue arises when *defining* the prior beliefs that justify a randomized trial with equal probabilities of control and treatment in the first place. A detour into the famous ECMO case in medical ethics is warranted.

Bartlett et al. [1982] reported results from "phase I" trials of safety, side effects and effects of variants on procedures. These trials had been conducted over 8 years, and interim results reported in 1977 and 1980. They found that 25 of 45 patients survived, and most with no side effects of note. They also reported (p. 429) that other clinics had adopted their surgical procedures for comparable cases, and reported 15 of 23 survivors. In Bartlett et al. [1985; p. 479] they reported a larger sample of 55 from these phase I trials, noting a survival rate with ECMO of 70% for 40 of these infants with birth weight in excess of 2kg. Survival weights for infants born 2kg or less was much lower, resulting in an overall mortality rate of 56% for the 55 observational patients. In a comment, Ware and Epstein [1985; p. 851] note that these results "provide encouraging evidence for the efficacy of ECMO," which presumably translates into a subjective belief that ECMO is a superior treatment to some degree.

All of this is preparatory, but in the sequel critical, to the primary focus: sequential trials reported in Bartlett et al. [1985] and then O'Rourke et al. [1989]. Ware [1989] reviewed the statistical rationale for the experimental design of the trial by O'Rourke et al. [1989], and encounters a blistering tsunami of critical commentary.

[20] I never look to the processes or decisions of institutional review boards for guidance on ethical matters, although certainly respect the idea of informed consent and the fact that one should have a "gatekeeper" for their approval. I have even less patience for the illusion that pre-registries mitigate poor scholarship.

The trial of Bartlett et al. [1985] used a randomized play-the-winner rule for deciding which treatment, ECMO or Conventional Medical Therapy (CMT), would be used for the next patient. Each treatment started with one ball each in an urn, and one was selected at random. Every time a selected treatment led to survival, an extra ball for that treatment was added to the urn. And every time a selected treatment led to a death, an extra ball was added for the *other* treatment. The first patient was randomized to ECMO and survived, so coming into the allocation of treatment for the second patient the chance of being allocated ECMO was 0.67 = 2÷3. The second patient was allocated to CMT and died, so coming into the allocation of treatment for the third patient the chance of being allocated ECMO was 0.75 = 3÷4. A stopping rule had been set by which the trial would be terminated when 10 CMT patients had died or 10 patients had been allocated to ECMO. As it happens, the third and all subsequent patients were allocated to ECMO and survived. Two additional patients met the selection criteria, were randomly assigned to ECMO, and survived. Hence, in the end, there were 12 patients: all 11 assigned to ECMO survived and the 1 assigned to CMT died.

The trial of O'Rourke et al. [1989] was motivated primarily by a concern with the fact that only one CMT patient was evaluated by Bartlett et al. [1985]. This issue was raised by Ware and Epstein [1985] in a commentary on Bartlett et al. [1985], explicitly stressing the

> ... conflict between the interests of the individual patient and the interests of the population of similar future patients whose care will be influenced by the results of the trial. Medical ethics requires that physicians give primary consideration to the well-being of the individual patient under their care; yet, most clinical trials are designed to continue to a fixed sample size, even if interim results strongly suggest (but do not prove) the superiority of oine oif the regimens under study. This strategy is employed in the belief that it will yield maximum benefit to the patiemnt population as a whole and is usually justified to both caretakers and investigators by the arhument that neither treatment has been shown to be superior. As a substantial trend in favoir of one of the regimens emerges, hwoever, a potential conflict arises between the desire to choose the more promising therapy for the next patient and the need to gather additional comparative information. (p.850)

It is also worth being reminded, by Ware [1989; p. 300], that

> This period was one of intense debate between proponents of ECMO, who believed that the therapy was a breakthrough in treatment [...], and skeptics, who were unconvinced by the registry data on mortality rates and expressed concerns about potential morbidity of ECMO treatment, especially brain hemorrhage and subsequent severe impairment.

Furthermore, Begg [1989; p. 320] also reminds us that the boring issue of "covariate balance" looms large with small samples:

> A more serious problem, however, is the potential for covariate imbalance between the treatment groups. In large studies, we can be confident that randomization distributes the poor risk and good risk patients in an evenhanded way. However, in small studies [...], serious covariate imbalance is quite likely and may well explain unusual results.

Begg [1989] concluded the trial of O'Rourke et al. [1989] was stopped too *early*, since the resulting inferences were not convincing to a wider audience. In fact, it is telling that even after ECMO was introduced into the U.K. in 1989, it was deemed necessary to undertake a massive clinical trial (Fields et al [1996]). So these are not simple statistical issues to be resolved crisply.

Nonetheless, Royall [1989] and Berry [1989; p. 306] reject the claim that prior evidence from the randomized evaluation documented by Bartlett et al. [1985] supported such a perfectly diffuse prior. Royall [1989; p. 318] calculates the posterior probability that the ECMO treatment was inferior to be either 0.01 or 0.00003 based on previous data. Kass and Greenhouse [1989; p. 313] raise similar concerns, but in the end explicitly, and reluctantly, just *assume* that the study was "appropriately designed" to start with a diffuse prior. Hence they defer to the judgment of those that designed the experiment, on the grounds that they would not have ethically started the experiment with a diffuse prior unless they *actually adopted* a position of "clinical equipoise" with respect to the treatments.[21]

Berry [1989; p.310] sharply concludes that "clinical equipoise is an invention used to avoid difficult ethical questions." In the context of economics experiments, that equipoise corresponds to

[21] To economists: this claim should not be read on an "as if" basis, literally and also methodologically.

claims that "anything *could* happen," as distinct from "here is what I believe *would* happen." Freedman [1987] first proposed the notion of clinical equipoise, controversially defining it in terms of priors that are presumed to be held in the broader research field, not the priors of the immediate investigators.

In general we need to be able to pool disparate sources of data, even observational studies, to form priors for ethical grounds *prior* to randomization, and that type of pooling is exactly what Bayesian analysis facilitates. Setting aside for the moment the claims by some that the previous non-experimental and experimental data for ECMO was sufficient never to have started further experiments, the general point is only that we should not default to diffuse priors unless there are strong grounds for doing so.

B. During and After the Experiment

The ethical need for being able to systematically pool priors and data also arises *during and after* the trial, when determining what to make of the results in the context of many other sources of information that are *not* directly comparable (i.e., exchangeable). This issue arises so often that it cannot be set aside from the instant trial.[22] An example in Section 4 illustrates this point for economists.

The other ethical issue that arises during the experiment is the consenting process. In the ECMO case another controversy arose with parents randomly selected to have their baby receive the CMT *not being asked for consent*. The argument is that CMT was the default anyway, and that ECMO was the "unproven" treatment that required consenting. Concerns about the distress that parents must undergo for any consenting process should, of course, be respected. But the asymmetry of the concern is problematic, particularly in the absence of clinical equipoise. As the evidence in favor of ECMO mounted, this asymmetry in consenting raised even more concerns.[23]

[22] See Peto [1985; p. 33] and Armitage [1985; p.19/20] for discussion in the context of medical trials.

[23] Without knowing the full institutional details of the ECMO trials, it is common, but not universal, for a general consenting process to explain this issue, and that priors in favor of the alternatives were diffuse

In some settings in economics there is no informed consent at all, such as in the "natural field experiment" defined by Harrison and List [2004] as part of a taxonomy of field experiments. The signature characteristic of these experiments is that the subject does not know that they are in an experiment. Often these just reflect government policies that contain random elements or discontinuities, and the sole role of the experimenter is to evaluate the data provided by the policy as if it had been designed *ex ante* as an experiment. One of the most elegant examples of a natural field experiment designed *ex ante* to avoid consenting is due to Camerer [1998], who simply placed certain bets at a race track to examine if asset markets can be manipulated.[24] Defenses for a complete lack of consenting can be tenuous, but exist.[25]

to start off with. These processes often explain that sequential outcomes from the trial have not been provided to the medical professionals undertaking it. Occasionally, there is even language about the dangers of making inferences from small samples as the rationale for wanting to complete the trial according to pre-set stopping rules.

[24] Camerer [1998] recognized that computerized betting systems allowed bets to be placed and cancelled before the race was run. Thus he could try to manipulate the market by placing bets in certain ways to move the market odds, and then cancelling them. The cancellation kept his net budget at zero, and in fact is one of the main treatments, to see if such a temporary bet changes prices appreciably. He found that it did not, but the methodological cleanliness of the test is remarkable.

[25] For example, List [2008; p.672] cites a field experiment conducted with "... a national fundraiser to explore various methods that fundraisers might wish to implement to be able to provide more of the public good. During the research, we never learned the solicitees' names, solicitees received letters similar to the ones they were sent in the normal course of their lives, and they made charitable donation decisions in a natural manner. In the end, we learned something interesting about the economics of charity while *doing no harm to the solicitees*. Indeed, *some might argue* that these potential donors were better off *because our methods induced more giving and therefore a higher provision of the public good*. When the research *makes participants better off, benefits society*, and confers anonymity and just treatment to all subjects, the lack of informed consent seems defensible. Ethical issues surrounding human experimentation are of utmost importance. Yet, the benefits and costs of informed consent should be carefully considered in each situation. Those cases in which there are *minimal benefits* of informed consent but *large costs* are prime candidates for relaxation of informed consent (emphasis added)" Although this is a relatively innocuous case, the lessons do not readily generalize. What is the metric from welfare economics for asserting that no harm was done to the subjects? Why would increased provision of public good X, and potentially reduced provision of public good Y in a zero-sum contribution setting, make this subject or society better off? Why does the argument *of some* observers in favor of there being benefits to participants become a *presumption* that the participants are better off? My concern is the general evidentiary basis of the judgment as to when the benefits are minimal and the costs are large.

5. Implications

Return to the worked example from Section 3.B, we illustrate how one can undertake *adaptive* welfare evaluation during an experiment.[26] Recall that the worked example involved subjects from Harrison and Ng [2016] making 24 decision to purchase insurance or not, where the expected CS from their decision took the form, illustrated in Figure 1, of posterior predictive distributions.

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? Important recent research by Hadad et al. [2020] and Kasy and Sautmann [2019] 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.

We consider a simple application of the Bayesian approach to behavioral welfare economics to illustrate some important issues. 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. The order of insurance products, differentiated by their actuarial parameters, was randomly assigned to each subject.[27] 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

[26] The exposition in this sub-section is adapted from Gao, Harrison and Tchernis [2020; §3.C].

[27] 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.

observed decision 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 decision. Since this is a distribution, we can evaluate the Bayesian probability that *each* decision 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.[28] 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 declines more or less steadily towards 0.5 over time, the probability of DNH is always greater than 50:50 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 probability of DNH as the probabilities of positive or negative CS outcomes, given the risk preferences of the subject. So the fact that the 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 can think of the *policy maker* as exhibiting risk neutral 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

[28] 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 has long been common to employ termination rules based on extreme, cumulative results (e.g., the "3 standard deviations" rule noted by Peto [1985; p. 33]).

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 much 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 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.[29] 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.[30]

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

[29] 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. Andersen, Harrison, Lau and Rutström [2008][2014] consider the joint estimation of risk and time preferences. Berry and Fristedt [1985; chapter 3] stress the importance of time discounting in sequential "bandit" problems in medical settings.

[30] The intertemporal risk aversion of a subject, also referred to as "correlation aversion," bears no necessary relationship to atemporal risk aversion. Andersen, Harrison, Lau and Rutström [2018] consider the joint estimation of atemporal risk preferences, time preferences, and intertemporal risk preferences.

classical utilitarian social welfare function over all 111 subjects.

One general lesson from this case study 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. And this lesson arises on top of the lessons from the ECMO case study about the normative basis of experimentation in the first place, based on prior evidence from non-experimental or controlled environments.

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 that experiment. If we had truncated these experiments adaptively as suggested, 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 field.

6. Conclusion

If we take seriously the intent to improve welfare for individuals with interventions, then we must allow that we are also capable of doing harm in terms of welfare. The normative and ethical issues that arise for experimental design and interpretation lead to a derived demand for more economic theory to be used, and rigorous econometric methods. Specifically, we need to work out how to do behavioral welfare economics in a rigorous and general manner, and we need to consistently apply Bayesian inference methods to avoid the application of dogmatic or *ad hoc* priors.

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

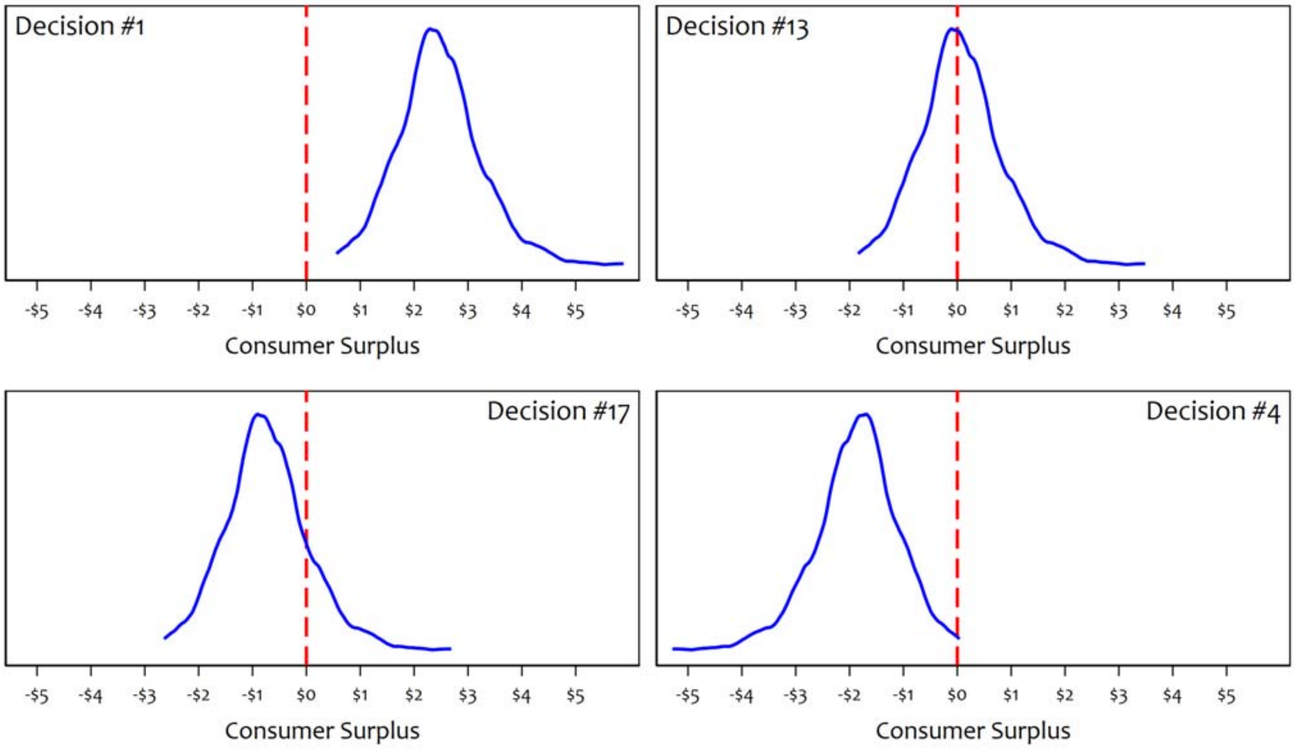


Figure 2: Adaptive Welfare Evaluations for Subject #1

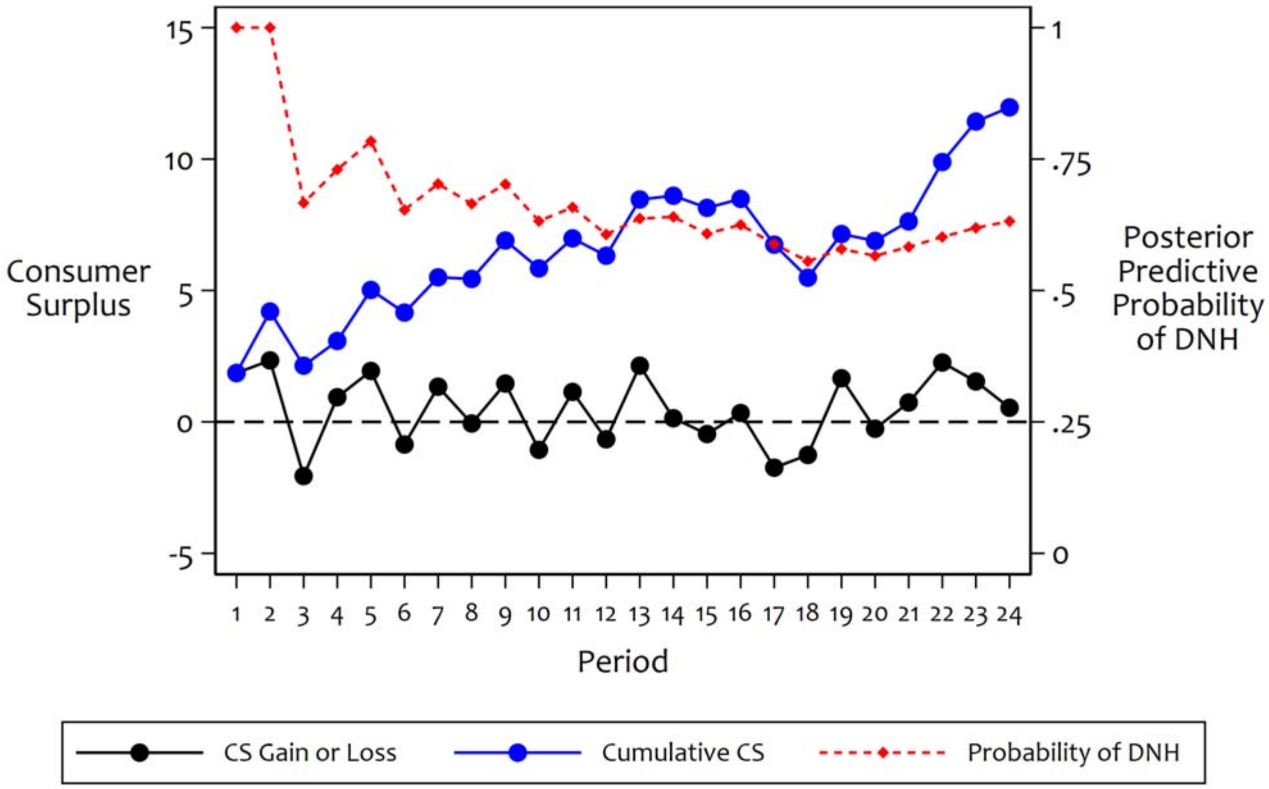
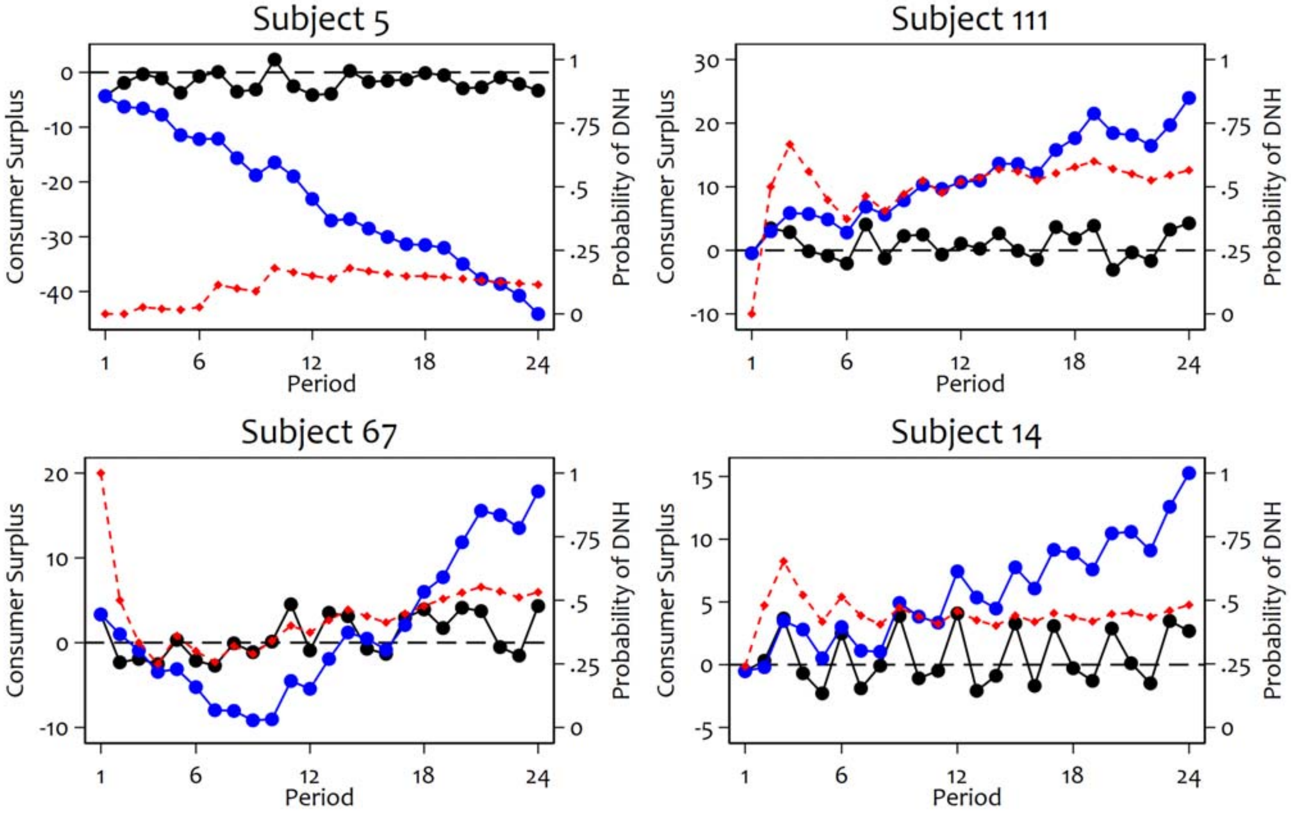


Figure 3: Individual Adaptive Welfare Evaluations for Four Subjects



Subject 5

Subject 111

Subject 67

Subject 14

References

Armitage, Paul, "The Search for Optimality in Clinical Trials," *International Statistical Review*, 53(1), 1985, 15-24.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten Igel, and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), May 2008, 583-618.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten Igel, and Rutström, E. Elisabet, "Discounting Behavior: A Reconsideration," *European Economic Review*, 71, November 2014, 15-33.

Andersen, Steffen; Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet, "Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion," *International Economic Review*, 59(2), May 2018, 537-555.

Bartlett, Robert H.; Andrews, Alice F.; Toomasian, John M.; Haidue, Nick J., and Gazzaniga, Alan B., "Extracorporeal Membrane Oxygenation for Newborn Respiratory Failure: Forty-Five Cases," *Surgery*, 92(2), 1982, 425-433.

Bartlett, Robert H.; Roloff, Dietrich W.; Cornell, Richard G.; Andrews, Alice French; Dillon, Peter W., and Zwischenberger, Joseph B., "Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study," *Pediatrics*, 76(4), October 1985, 479-487.

Bernheim, B. Douglas, "Behavioral Welfare Economics," *Journal of the European Economic Association*, 7(2–3), 2009, 267–319.

Bernheim, B. Douglas, "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics," *Journal of Benefit-Cost Analysis*, 7(1), 2016, 12–68.

Bernheim, B. Douglas, and Rangel, Antonio, "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics*, 124(1), 2009, 51–104.

Begg, Colin B., "Comment: Ethics and ECMO," *Statistical Science*, 4(4), 1989, 320-322.

Berry, Donald A., "Comment: Ethics and ECMO," *Statistical Science*, 4(4), 1989, 306-310.

Berry, Donald A., and Fristedt, Bert (eds.), *Bandit Problems: Sequential Allocation of Experiments* (New York: Springer, 1985).

Binmore, Ken, *Rational Decisions* (Princeton: Princeton University Press, 1999).

Bogdan, Radu J., *Interpreting Minds* (Cambridge, MA: MIT Press, 1997).

Bossuroy, Thomas, and Delavallade, Clara, "Experiments, Policy, and Theory in Development Economics," *Journal of Economic Methodology*, 23(2), 2016, 147-156.

Burge, Tyler, "Individualism and Psychology," *Philosophical Review*, 95(1), 1986, 3-45.

Camerer, Colin F., "Can Asset Markets Be Manipulated? A Field Experiment with Racetrack Betting," *Journal of Political Economy*, 106(3), 1998, 457-482.

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

Clark, Andy, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (New York: Oxford University Press, 2015).

Cox, James C.; Roberson, Bruce; and Smith, Vernon L., "Theory and Behavior of Single Object Auctions," in V.L. Smith (ed.), *Research in Experimental Economics* (Greenwich: JAI Press, volume 2, 1982).

Cox, James C.; Smith, Vernon L., and Walker, James M, "Theory and Behavior of Multiple Unit Discriminative Auctions," *Journal of Finance*, 39(4), 1984, 983-1010.

Darwin, Charles, *The Variation of Animals and Plants Under Domestication* (London: John Murray, First Edition, 1868).

Dennett, Daniel C., "Intentional Systems in Cognitive Ethology: the 'Panglossian Paradigm' Defended," *Behavioral and Brain Sciences*, 6, 1983, 343-390.

Dennett, Daniel C., *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

Dennett, Daniel C., *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W.W. Norton and Company, 2016).

Ferber, Robert, and Hirsch, Werner Z., "Social Experimentation and Economic Policy," *Journal of Economic Literature*, XVI, December 1978, 1379-1414.

Field, D.J.; Davis, C.; Elbourne, D.; Grant, A.; Johnson, A. and Macrae, D., "UK Collaborative Randomised Trial of Neonatal Extracorporeal Membrane Oxygenation," *The Lancet*, 348, 1996, 75-82.

Forsythe, Robert; Palfrey, Thomas R., and Plott, Charles R., "Futures Markets and Informational Efficiency: a Laboratory Examination," *Journal of Finance*, 39(4), 1984, 955-981.

Freedman, Benjamin, "Equipoise and the Ethics of Clinical Research," *New England Journal of Medicine*, 317(3), 1987, 141-145.

Friedman, Daniel; Harrison, Glenn W., and Salmon, Jon, "The Informational Efficiency of Experimental Asset Markets," *Journal of Political Economy*, 92, June 1984, 349-408.

Gao, Xiaoxue Sherry; Harrison, Glenn W., and Tchernis, Rusty, "Behavioral Welfare Economics and

Risk Preferences: A Bayesian Approach," *NBER Working Paper 27685*, National Bureau of Economic Research, 2020.

Gibson, James J., "The Theory of Affordances," in R.Shaw & J.Bransford (eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (Hillsdale, NJ: Lawrence Erlbaum, 1977).

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

Hadad, Vitor; Hirshberg, David A.; Zhan, Ruohan; Wager, Stefan, and Athey, Susan, "Confidence Intervals for Policy Evaluation in Adaptive Experiments, *Working Paper*, Stanford University, July 2020; available at https://arxiv.org/abs/1911.02768.

Harrison, Glenn W., "Randomisation and Its Discontents," *Journal of African Economies*, 20(4), 2011a, 626-652.

Harrison, Glenn W., "Experimental Methods and the Welfare Evaluation of Policy Lotteries," *European Review of Agricultural Economics*, 38(3), 2011b, 335-360.

Harrison, Glenn W., "Field Experiments and Methodological Intolerance," *Journal of Economic Methodology*, 20(2), 2013, 103-117.

Harrison, Glenn W., "Impact Evaluation and Welfare Evaluation," *European Journal of Development Research*, 26, 2014a, 39-45.

Harrison, Glenn W., "Cautionary Notes on the Use of Field Experiments to Address Policy Issues," *Oxford Review of Economic Policy*, 30(4), 2014b, 753-763.

Harrison, Glenn W., "Field Experiments and Methodological Intolerance: Reply," *Journal of Economic Methodology*, 23(2), 2016, 157-159.

Harrison, Glenn W., "The Behavioral Welfare Economics of Insurance," *Geneva Risk & Insurance Review*, 44(2), September 2019, 137–175.

Harrison, Glenn W., "Field Experiments and Public Policy: *Festina Lente*," *Behavioural Public Policy*, 1-8. doi:10.1017/bpp.2020.28, 2020.

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

Harrison, Glenn W., and Ng, Jia Min, "Evaluating the Expected Welfare Gain from Insurance," *Journal of Risk and Insurance*, 83(1), 2016, 91–120.

Harrison, Glenn W., and Ross, Don A. "Varieties of Paternalism and the Heterogeneity of Utility Structures," *Journal of Economic Methodology*, 25(1), 2018, 42-67.

Hausman, Daniel, *Preference, Value, Choice and Welfare* (Cambridge: Cambridge University Press, 2011).

Hey, John D., "Why We Should Not Be Silent About Noise," *Experimental Economics*, 8(4), December 2005, 325–345.

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

Hohwy, Jakob, *The Predictive Mind* (New York: Oxford University Press, 2013).

Infante, Gerardo; Lecouteux, Guilhem, and Sugden, Robert, "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioral Welfare Economics," *Journal of Economic Methodology*, 23, 2016, 1-25.

Johnson, George, *The Ten Most Beautiful Experiments* (New York: Knopf, 2008).

Kass, Robert E., and Greenhouse, Joel B., "Comment: A Bayesian Perspective," *Statistical Science*, 4(4), 1989, 310-317.

Kasy, Maximilian, and Sautmann, Anja, "Adaptive Treatment Assignment in Experiments for Policy Choice, *Working Paper*, Oxford University, December 2019; available at https://maxkasy.github.io/home/research/, *Econometrica*, forthcoming.

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

Leamer, Edward E., "Tantalus on the Road to Asymptotia," *Journal of Economic Perspectives*, 24(2), 2010, 31-46.

List, John A., "Informed Consent in Social Science," *Science*, 322, October 31, 2008, 672.

McClamrock, Ron, *Existential Cognition: Computational Minds in the World* (Chicago, IL: University of Chicago Press, 1995).

O'Rourke, P. Pearl; Crone, Robert K.; Vacanti, Jospeh P.; Ware, James H.; Lillehel, Craig W.; Parad, Richard B., and Epstein, Michael F., "Extracorporeal Membrane Oxygenation and Conventional Medical Therapy in Neonates With Persistent Pulmonary Hypertension of the Newborn: A Prospective Randomized Study," *Pediatrics*, 84(6), 1989, 957-963.

Peto, Richard, "Discussion of Papers by J.A. Bather and P. Armitage," *International Statistical Review*, 53(1), 1985, 31-34.

Ross, Don, *Philosophy of Economics* (London: Palgrave Macmillan, 2014).

Roth, Alvin E., and Malouf, Michael W., "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, 86(6), 1979, 574–594.

Royall, Richard, "Comment," *Statistical Science*, 4(4), 1989, 318-319.

Seyfarth, Robert M., and Cheney, Dorothy L., "Dennett's Contribution to Research on the Animal Mind, " in A. Brook and D. Ross (eds.) *Daniel Dennett* (New York: Cambridge University Press, 2002).

Silverman, William A., *Retrolental Fibroplasia: A Modern Parable* (New York: Grune & Stratton, 1980); available at http://www.neonatology.org/classics/parable/.

Smith, Vernon L., "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy*, 70(2), 1962, 111-137.

Sugden, Robert, *The Community of Advantage: A Behavioural Economist's Defence of the Market* (New York: Oxford University Press, 2018).

Teele, Dawn Langan, "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* (New Haven, NJ: Yale University Press, 2014).

Ware, James H., "Investigating Therapies of Potentially Great Benefit: ECMO," *Statistical Science*, 4(4), 1989, 298-306.

Ware, James H., and Epstein, Michael F., "Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study – Comment," *Pediatrics*, 76(5), 1985, 849-851.

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

Wilcox, Nathaniel T., "Robert A. Millikan Meets the Credibility Revolution," *Journal of Economic Methodology*, 23(2), 2016, 130-138.

Worrall, John, "Why There's No Cause to Randomize," *British Journal of the Philosophy of Science*, 58, 2007, 451–488.