Modeling the Modeler: A Normative Theory of Experimental Design*

Fernando Payró† and Evan Piermont‡

January 30, 2025

Abstract

We consider an analyst whose goal is to identify a subject's utility function through revealed preference analysis. We argue the analyst's preference about which experiments to run should adhere to three normative principles: The first, *Structural Invariance*, requires that the value of a choice experiment only depends on what the experiment may potentially reveal. The second, *Identification Separability*, demands that the value of identification is independent of what would have been counterfactually identified had the subject had a different utility. Finally, *Information Monotonicity* asks that more informative experiments are preferred. We provide a representation theorem, showing that these three principles characterize *Expected Identification Value* maximization, a functional form that unifies several theories of experimental design. We also study several special cases and discuss potential applications.

Keywords: Experimental Design, Partial Identification, Revealed Preferences

JEL Classification: D81

Declarations of interest: None

*We thank seminar participants at Bart Lipman's Conference, Autonomous University of Barcelona and Centre d'Économie de la Sorbonne.

†Autonomous University of Barcelona, Barcelona School of Economics and Center for the Study of Organizations and Decisions in Economics; fernando.payro@uab.cat. Payró gratefully acknowledge financial support from the Ministerio de Economía y Competitividad and Feder (PID2020-116771GB-I00 and PID2023-147183NB-I00) and financial support from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S)

‡Royal Holloway, University of London, Department of Economics; evan.piermont@rhul.ac.uk.

1 Introduction

This paper proposes a theory of experimental design. We consider an analyst who's goal is to identify a subject's utility function through the use of experiments, specifically, by offering the subject decision problems and observing his choices. Because of time, cost, or computational constraints, the analyst will only be able to offer a limited set of decision problems; how then should she choose which experiments to conduct? In this paper, we advance three normative principles and argue that they should guide any rational experimental design, independent of the specific objectives of the analyst.

As we explain in detail below, each observation in an experiment (partially) identifies some set of utilities, those that that are consistent with the observation. For example, observing a subject known to be a CRRA utility maximizer reject an actuarially fair lottery would identify him as risk averse, but might not resolve anything further about their coefficient of risk aversion. With this notion of partial identification in mind, the three normative principles can be stated as follows: First, *Information Monotonicity*, asserts that the analyst prefers sharper identification; that is, if one experiment always identifies a smaller subset of utilities than another experiment, it is preferred. Second, *Structural Invariance*, maintains that the value of an experiment should depend only on what it allows the analyst to identify and not on other structural details. Specifically, if two experiments yield the same set of possible inferences about the subject, they must be valued equally. Finally, *Identification Separability* demands that the value of making some partial identification should depend only on what was identified, and not on what the experiment would have counterfactually identified had the subject made a different choice.

Our main result shows that an analyst's preference over experiments adheres to these principles if and only if she seeks to maximize *expected identification value*, as we now explain. A rational analyst should be able to assign a value to each partial identification, that is, a value for learning that the subject's true utility lies is some subset. For example, an analyst can reflect on the relative value of learning a CRRA subject "is risk averse" versus learning the subject "has a coefficient of risk aversion in $[1.5, 2]$." This value of identification is subjective and encodes the analyst's

particular goals and motivations.[1]

Given a ex-ante distribution over the utility type of the subject, interpreted as the analyst's prior belief about the subject's preferences, each experiment will induce a distribution over consequent partial identifications: our three normative principles characterize ranking experiments in accordance with the expected value of the identification they permit. By accommodating any subjective value of identification, our theory unveils the common facets of rational experimental design that are independent of the idiosyncratic objectives of the analyst. Indeed, we detail how our approach unifies several distinct experimental paradigms. We then show how this theory provides additional insight in specific environments, where the normative principles settle concrete design choices. In particular, we specialize the model in two ways. First, by examining the case where the subject is known to be an expected utility maximizer. Second, by examining the case where the analyst seeks to maximize the reduction in entropy between her prior and posterior.

Discussion of Model and Results • To keep the analysis simple, we abstract away from the physical details of an experiment. Instead, we model an experiment as a menu of alternatives A and a partition of the menu $\mathscr{P}$ that captures what is observable to the analyst; when the subject chooses $a \in A$, the analyst observes the (unique) $P \in \mathscr{P}$ such that $a \in P$. When $\mathscr{P}$ is the discrete partition, the subject's choice is perfectly observed, as is likely in very simple environments, e.g., single-stage laboratory experiments. By allowing $\mathscr{P}$ to be coarser, we allow for more complex experimental environments. For example, A could represent the set of strategies in a dynamic environment where $\mathscr{P}$ captures the unobservability of off-path behavior.[2] In less controlled environments, observational restrictions are common-place, e.g., an online platform (google) might be able to observe which retailer was chosen by a user, but not the user's actual purchase. We take as primitive a preference over *randomized experiments*, that is, finitely supported distributions π over experiments. In the literature, such experiments are called discrete choice experiments. We also take as part of our primitive the set of ex-ante possible utility functions from which the

[1] For instance, an analyst may only be only interested in classifying subjects as risk averse or not, but uninterested in any further details of the subject's preference. Such an analyst would be indifferent between the two partial identifications from the prior sentence.

[2] This interpretation is discussed in Section 7.

subject's true utility, u^*, was drawn and a probability μ over $\mathcal{U}$, capturing the prior beliefs of the analyst.[3]

Given an experiment $(A, \mathscr{P})$ and an observation $P \in \mathscr{P}$, let $W_{A,P} \subseteq \mathcal{U}$ denote those utility functions that would choose an element in P out of A. Thus, the observation that an element of P was chosen out of A induces the partial identification of $W_{A,P}$, i.e., the analyst infers that $u^* \in W_{A,P}$. Our main result shows that Structural Invariance, Identification Separability, and Information Monotonicity[4] hold if and only if the analyst assigns a value, $\tau(W)$, to each possible partial identification, $W \subseteq \mathcal{U}$, and values experiments according to the expected value of the identification they will yield. Formally, the analyst's preference must be representable by a *expected identification value* functional of the form

$$F(\pi) = \sum_{\text{supp}(\pi)} \Big(\sum_{P \in \mathscr{P}} \tau(W_{A,P}) \mu(W_{A,P}) \Big) \pi(A, \mathscr{P}) \qquad (\star)$$

where π is a lottery over experiments, τ is interpreted as an *identification index* and μ as the analysts prior.

Since τ depends only on what can be inferred from the observed outcome, each experiment is equated with the sets of utilities it can partially identify. This reflects Structural Invariance. Moreover, the representation is additive across cells of the partitions and thus, the ranking between two experiments is independent of whatever they commonly identify. This reflects Identification Separability. Finally, Information Monotonicity, requires the identification index τ to always find information weakly beneficial: for all disjoint $W, W' \subseteq \mathcal{U}$ (set $V = W \cup W'$) it must be that

$$\tau(W)\mu(W|V) + \tau(W')\mu(W'|V) \geq \tau(V)$$

where $\mu(\cdot|V)$ is the conditional of μ given V. That is, given that the analyst can already identify V, the expected value of further learning the distinction between

[3] As this is a normative exercise, we are interested in providing a potential experimenter with guidance on how to construct rational preferences, rather than identifying her prior beliefs; as such we take μ as part of the primitive. Our theory can be applied, modulo certain technicalities, in the event there is no prior beliefs, as discussed in Section 6.

[4] Along with an an axiom dictating that the analyst is an expected utility maximizer with respect to the randomization across experiments.

W and W' is weakly positive.

Normative Principles as Guiding Design Choices • To better understand how Structural Invariance and Identification Separability relate to experimental design choices, we now consider two simple examples within the environment of expected utility maximizing subjects. We apply our general theory to this setting in Section 4.

In particular, assume that the subject entertains a linear utility function over lotteries defined on three alternatives $\{a, b, c\}$. Denote by $\alpha x + (1-\alpha)y$ the lottery that places probability α on alternative x and $(1-\alpha)$ on y, for $x, y \in \{a, b, c\}$. Consider an analyst who needs to choose one of the following two experimental procedures. Both experimental procedures offer two menus to the subject:

$$\begin{aligned}
\textsc{exp a}: A &= \{a,\ \tfrac{1}{2}a + \tfrac{1}{2}b,\ \tfrac{1}{2}a + \tfrac{1}{2}c,\ \tfrac{1}{2}b + \tfrac{1}{2}c\} \\
A' &= \{\tfrac{6}{10}b + \tfrac{4}{10}c,\ \tfrac{4}{10}b + \tfrac{6}{10}c\}. \\
\textsc{exp b}: B &= \{a,\ b,\ c\} \\
B' &= \{\tfrac{2}{3}a + \tfrac{1}{3}b,\ \tfrac{2}{3}a + \tfrac{1}{3}c,\ \tfrac{1}{3}a + \tfrac{1}{3}b + \tfrac{1}{3}c\}.
\end{aligned}$$

These are visualized in the top of Figure 1.

Which of these two experiments should the analyst run? At first glance, this appears to be a matter of taste, as it seems plausible the answer should depend on the analyst's objectives, that is, on which aspects of the subject's preference she is interested in identifying. However, the principle of Structural Invariance imposes that these two experiments must be valued equally, as they induce the same possible set of partial identifications. To see this, observe that because expected utility is linear, observing a choice from A and A' is informationally equivalent to observing a single choice from $\{\frac{1}{2}x + \frac{1}{2}x' \mid x \in A, x' \in A'\} \equiv \frac{1}{2}A + \frac{1}{2}A'$ (and likewise for B and B'). Moreover, as shown in the bottom of Figure 1, $\frac{1}{2}A + \frac{1}{2}A'$ and $\frac{1}{2}B + \frac{1}{2}B'$ yield the same identifiable sets.[5]

Within the domain of linear utility, Structural Invariance is captured by a translation invariance axiom, which arises from the specific characteristics of the en-

[5]For example, the set of utilities which find $\frac{1}{2}(\frac{1}{2}a + \frac{1}{2}b) + \frac{1}{2}(\frac{6}{10}b + \frac{4}{10}c)$ maximal from $\frac{1}{2}A + \frac{1}{2}A'$ is exactly those that find $\frac{1}{2}b + \frac{1}{2}(\frac{2}{3}a + \frac{1}{3}b)$ maximal from $\frac{1}{2}B + \frac{1}{2}B'$ (the set W_1).

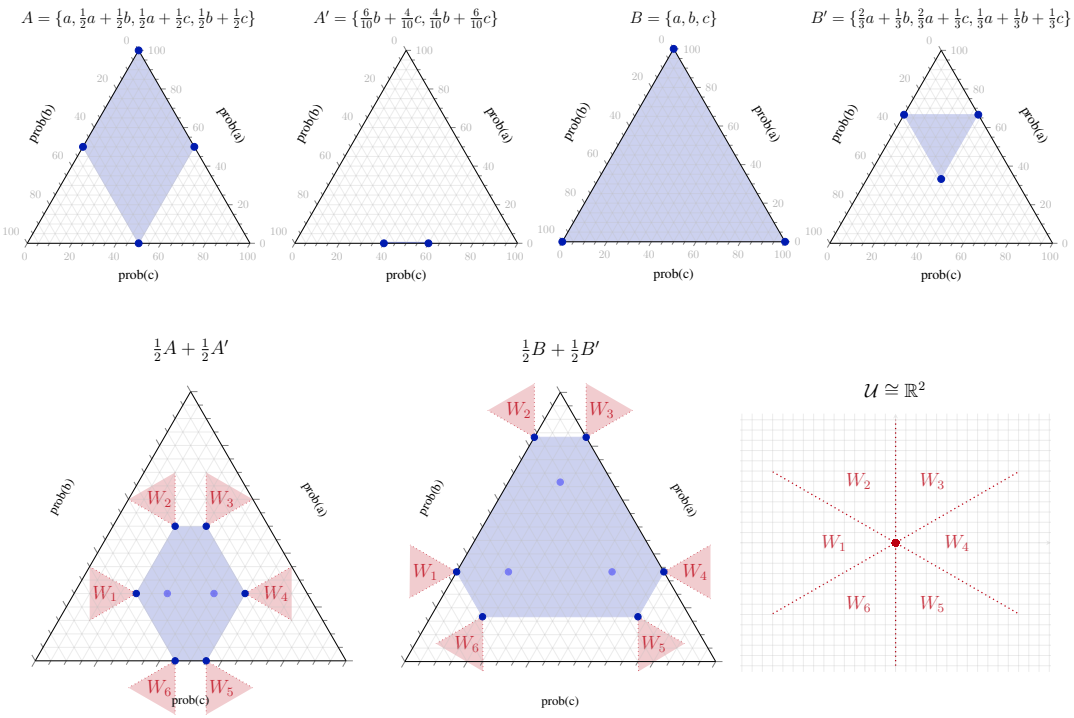

Figure 1: *Top*: The four decision problems in experiments EXP A and EXP B represented in the simplex. The convex hull of the decision problems is shaded. *Bottom*: The identifiable sets from $\frac{1}{2}A + \frac{1}{2}A'$ and $\frac{1}{2}B + \frac{1}{2}B'$. These form the same partition of $\mathcal{U}$ as shown in the third panel.

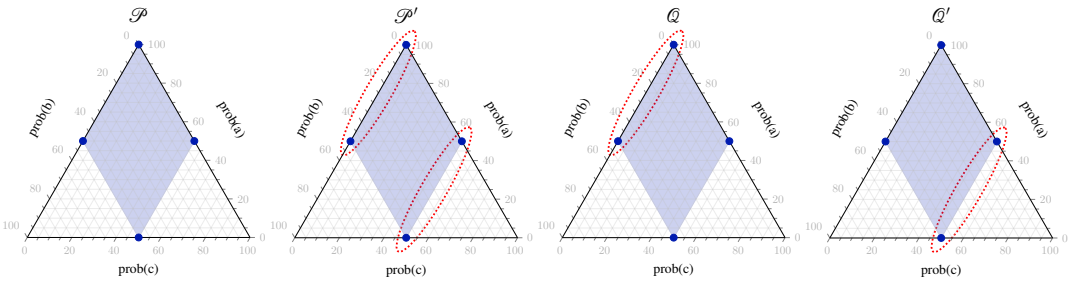

Figure 2: The four partitions of A from EXP C and EXP D.

vironment. Conceptually, Structural Invariance states that changing aspects of an experiment that do not affect what it can identify should not change its value; linear translations are the specific structural property invariant for linear utilities. As such, applying our results for expected utility (Section 4) to non-linear models (e.g. ambiguity averse utility functions) only requires identifying the appropriate invariance axiom.

To understand Identification Separability, consider the following four partitions of the decision problem A (from the earlier example); these partitions are shown in Figure 2.

$$\begin{aligned}
\mathscr{P} &= \{\{a\},\ \{\tfrac{1}{2}a + \tfrac{1}{2}b\},\ \{\tfrac{1}{2}a + \tfrac{1}{2}c\}, \{\tfrac{1}{2}b + \tfrac{1}{2}c\}\} \\
\mathscr{P}' &= \{\{a,\ \tfrac{1}{2}a + \tfrac{1}{2}b\},\ \{\tfrac{1}{2}a + \tfrac{1}{2}c, \tfrac{1}{2}b + \tfrac{1}{2}c\}\} \\
\mathscr{Q} &= \{\{a,\ \tfrac{1}{2}a + \tfrac{1}{2}b\},\ \{\tfrac{1}{2}a + \tfrac{1}{2}c\}, \{\tfrac{1}{2}b + \tfrac{1}{2}c\}\} \\
\mathscr{Q}' &= \{\{a\},\ \{\tfrac{1}{2}a + \tfrac{1}{2}b\},\ \{\tfrac{1}{2}a + \tfrac{1}{2}c,\ \tfrac{1}{2}b + \tfrac{1}{2}c\}\}
\end{aligned}$$

Based on these decision problems, the analyst considers two randomized experiments:

EXP C : The analyst offers $(A, \mathscr{P})$ and $(A, \mathscr{P}')$ each with probability with $\frac{1}{2}$.

EXP D : The analyst offers $(A, \mathscr{Q})$ and $(A, \mathscr{Q}')$ each with probability with $\frac{1}{2}$.

As in the previous example, what might seem to depend on the aims of the analyst is in fact dictated by criteria of rational design; the principle of Identification Separability requires these two experiments are valued equally. To see this, notice that

under EXP C, with probability $\frac{1}{2}$ the analyst learns exactly which element was chosen, and with the remaining probability $\frac{1}{2}$, she learns only which cell of $\mathscr{P}'$ contained the chosen alternative. Although less immediate, this is also the case for EXP D. Indeed, let $W \subset \mathcal{U}$ denote the set of utilities such that the subject will choose an element of the first cell of $\mathscr{P}'$ (i.e., will choose either a or $\frac{1}{2}a + \frac{1}{2}b$) and $W^c \subset \mathcal{U}$ those utilities such that the subject will choose an element of the second (i.e., either $\frac{1}{2}a + \frac{1}{2}c$ or $\frac{1}{2}b + \frac{1}{2}c$).[6] Then, conditional on $u^* \in W$, the subject will choose an element out of the first cell of $\mathscr{P}'$: under $\mathscr{Q}$ this is all that is observed, while under $\mathscr{Q}'$ the choice is observed perfectly. Conditional on $u^* \in W^c$, the same logic applies: $\mathscr{Q}$ perfectly reveals the subject's choice, while $\mathscr{Q}'$ only that the second cell of $\mathscr{P}'$ contains the chosen alternative.

Thus, both EXP C and EXP D reveal the subject's choice half of the time and the cell of $\mathscr{P}'$ containing his choice the other half. The difference between these two experiments is that in the latter, the amount of information revealed is correlated with the the subject's utility type. That is, if EXP D ends up perfectly revealing the subject's choice when $u^* \in W$, we know that it *would not have done* had $u^* \in W^c$, and vice versa. The principle of Identification Separability dictates that such counterfactual assessments are irrelevant, and thus, that the two randomized experiments are valued equally.

Functional Forms • The expected identification value representation is flexible enough to accommodate many Bayesian theories of optimal experimental design. For instance, by taking

$$\tau(W) = -\log(\mu(W)),$$

the value of an experiment is its expected reduction in entropy relative to the prior (Cover et al., 1991). We axiomatize this special case in Section 5. Drake et al. (2022) propose a dynamic Bayesian procedure for preference identification on the basis of this functional form. Another special case of our index comes from hypothesis testing. An analyst who wishes to test if the subjects preference is in some set W^*

[6] Using the notation from the bottom of Figure 1, $W = W_1 \cup W_2 \cup W_3$ and $W^c = W_4 \cup W_4 \cup W_5$; notice we are excluding the possibility of the subject being indifferent between alternatives. Following the literature on random utility, we assume such ties occur with zero probability. See section 2.

would consider

$$\tau(W) = \begin{cases} 1 & \text{if } W \subseteq W^* \text{ or } W^* \subseteq W^c \\ 0 & \text{otherwise .} \end{cases}$$

Finally, we observe that although our theory is meant to contribute to the conceptual understanding of choice experiments, it is flexible enough to allow for functional forms unrelated to experimental design. For instance, a Bayesian principal may only want to promote agents with similar preferences to herself. Hence, she conducts a test to see what kind of preferences her agents have when making promotion decisions. The following specification allows for such interpretation

$$\tau(W) = \max_{a \in \{0,1\}} \int_W \xi(a, u) d\mu.$$

where $a = 1$ ($a = 0$) is interpreted as (not) promoting the agent and $\xi(a, u)$ is her utility of promoting an agent that has preference u and μ is her prior over the agent's preference.

Outline • The paper proceeds as follows: The introduction concludes with a review of the relevant literature. The model is presented in Section 2. Our normative principles and main representation result are in Section 3. Sections 4 and 5 discuss the special cases of Expected Utility maximizing subjects and entropy reduction, respectively. Section 6 outlines a version of the model without prior beliefs. Finally, Section 7 concludes by showing how our framework is general enough to capture dynamic experiments. All proofs are collected in the Appendix.

1.1 Related Literature

This paper joins the large literature in economics on eliciting preferences from observable behavior. It differs from most of the literature as it does not focus on efficiency of a particular elicitation method, but on what are the minimal properties a method should satisfy in order to be considered rational. Such questions have been suggested in the statistics literature on Bayesian experimental design. Early texts such as Raiffa and Schlaifer (2000) and Lindley (1972) propose a utility function for Bayesian experimenters. The literature that followed provided specializations of

the general function to feet more stylized settings such as regression analysis and model discrimination analysis (see Chaloner and Verdinelli (1995) for a review of the literature). Thus, our work complements the existing literature by providing a framework to discuss experimental design as well as the missing axiomatic foundation.

Our framework is inspired by the literature in economics on Discrete Choice Experiments (henceforth DCE). DCE's were initially developed by Louviere and Hensher (1982). They are based on the theoretical framework of the Random Utility Model (Luce (1959) and McFadden (1973)). We contribute to the DCE literature by providing a unifying framework to analyse deviations from the standard DCE method. Given the recent interest in employing dynamic procedures to substitute DCE's, our results can be used as a test for such procedures. If they do not satisfy our axioms, they should not be employed.

Outside of the DCE literature but within the random utility literature, Gul and Pesendorfer (2006) (henceforth GP) provide necessary and sufficient conditions for random choice data to be consistent with *random* expected utility. We use their work as a building block in providing foundations to Bayesian procedures. Specifically, our richness condition described in Section 4 are direct consequences of the GP assumptions.

Gilboa and Lehrer (1991) studies a related problem to ours. Their goal is to provide axiomatic foundations for functions over partitions of states that can be interpreted as describing the value of information for some Bayesian agent. Our analysis can also be interpreted as providing foundations for functions that can be interpreted as describing the value of additional information of an agent's preference. There are two key differences. First, we take as observable preference over experiments as opposed to a function over partitions of the utility space, which would be the analog of their domain in our setting. Second, we do not look for identification functions for which there exists a Bayesian experimenter that may employ them. Indeed, we do not take rationality as given and look for identification functions that satisfy it. We propose a notion of rationality and characterize the set of identification functions that satisfy it.

Finally, our work also contributes to the preference over menus in decision the-

ory. Starting with Gul and Pesendorfer (2001) and Dekel et al. (2001), economists have used preferences over menus of lotteries to study distinct phychological phenomena such as temptation and self-control. The literature then generalized the domain to lotteries over menus of lotteries to obtain sharper results (Stovall (2018) and Ergin and Sarver (2015)). Our work shows that lotteries over menus can also be employed to analyze experimental design. Thus, it suggests that some of the earlier work in decision theory may be useful for experimental design.

2 General Model

2.1 Preliminaries

An *abstract experimental environment* is a tuple $(Z, \mathcal{U}, \Omega, \mu)$ where Z is some set of possible choice alternatives, $\mathcal{U} \subseteq \{u : Z \to R\}$ a set of utility types, Ω is an algebra of measurable sets of $\mathcal{U}$ and μ probability distribution over $(\mathcal{U}, \Omega)$. We interpret $(Z, \mathcal{U}, \Omega, \mu)$ as the theory the analyst has about the subject's preferences. A *decision problem* A is a finite subset of Z. Let $\mathbb{D}$ denote the set of all decision problems.

Given a decision problem A and some $B \subseteq A$, let

$$W_{A,B} = \{u \in \mathcal{U}, B \cap \arg\max_{x \in A} u(x) \neq \varnothing\}$$

denote the set of utilities for which some element of B is a maximizer when facing the decision problem A. Intuitively, $W_{A,B}$ is the set of preferences that would choose an element in B when facing menu A.

To achieve her goal, the analyst can can offer the subject a decision problem from which she will observe the subject's choice. While it is plausible that a subject's behavior can be observed perfectly in static laboratory conditions, in dynamic settings or more general environments (i.e., field experiments, consumer testing in industry, data collection by online platforms, etc.), the analyst may only be able to partially observe choice. To allow for such constraints, we define an experiment as a

decision problem and a partition.[7] The interpretation is that $P \in \mathscr{P}$ represents what the analyst observes when the subject's choice out of A is contained in P.

Formally, an *experiment* $e = (A, \mathscr{P})$ is a pair where where $A \in \mathbb{D}$ and $\mathscr{P}$ is a partition of A such that for any $P, Q \in \mathscr{P}$

(E1) $W_{A,P} \in \Omega$
(E2) $\mu(W_{A,P} \cap W_{A,Q}) = 0$

The first requirement states that analyst assigns a prior probability to each observable outcome, and the second states that the analyst can unambiguously interpret the observed outcome. Specifically, the analyst places μ-probability 0 on the subject being indifferent between two alternatives in A so that observed choice can be interpreted without worrying about how ties are broken.

Our notion of experiments can be used to define partial identification: A set of preferences $W \subseteq \mathcal{U}$ is *identifiable* in $(A, \mathscr{P})$ if $W = W_{A,P}$ for some $P \in \mathscr{P}$. Given an experiment $(A, \mathscr{P})$, the analyst can calculate the family $\{W_{A,P} | P \in \mathscr{P}\}$, the sets of preferences that are identifiable by the experiment.

Call two (finite) collections of subsets of $\mathcal{U}$, $\{W_1, \dots W_n\}$ and $\{V_1, \dots W_m\}$ μ-*equivalent* if for every W_i with $\mu(W_i) > 0$ there exists a V_j such that $\mu(W_i) = \mu(W_i \cap V_j) = \mu(V_j)$, and vice versa. That is, the collections are μ-equivalent if they are equal up to measure 0 sets.

We assume the analyst has access to a set of experiments $\mathbb{E}$ that satisfies the following two properties:

- $(A, \mathscr{P}) \in \mathbb{E}$ and $\mathscr{Q}$ is a coarsening of $\mathscr{P}$ then $(A, \mathscr{Q}) \in \mathbb{E}$,
- For any finite Ω-measurable partition $\mathcal{W}$ of $\mathcal{U}$, there is a some experiment $(A, \mathscr{P}) \in \mathbb{E}$ such that $\{W_{A,P} | P \in \mathscr{P}\}$ is μ-equivalent to $\mathcal{W}$.

The first property states that if $(A, \mathscr{P})$ is feasible, then an experiment that potentially identifies less utilities is also feasible. The second property demands that for any finite partition of the utility space, the analyst can always find an experiment that would induce such a partition. We call such sets of experiments *rich*.

[7] A partition $\mathscr{P}$ of X is a set of subsets of X that are mutually disjoint and whose union is X.

Given a set of experiments $\mathbb{E}$, a *randomized experiment* (over $\mathbb{E}$) π is a finitely supported probability distribution over $\mathbb{E}$. The set of all randomized experiments over $\mathbb{E}$ is denoted by $\Pi(\mathbb{E})$. For a given randomized experiment, π, let $\text{supp}(\pi) = \{e \in \mathbb{E} \mid \pi(e) > 0\}$ denote the support of π. Our primitive is the analyst's preference, $\succcurlyeq$, over the set of all randomized experiments over some rich set of experiments.

2.2 Representation

A *expected identification value representation* for $\succcurlyeq$ is the following:

$$F(\pi) = \sum_{\text{supp}(\pi)} \Big(\sum_{P \in \mathscr{P}} \tau(W_{A,P}) \mu(W_{A,P}) \Big) \pi(A, \mathscr{P}) \qquad (\star\star)$$

where $\tau : \Omega \to \mathbb{R}$ satisfies

(T1) For all non-μ-null V and $W \subseteq V$,

$$\tau(W)\mu(W \mid V) + \tau(V \setminus W)(1 - \mu(W \mid V)) \geq \tau(V),$$

with equality holding whenever $\mu(W) = 0$.

(T2) $\tau(\mathcal{U}) = 0$.

Condition (T1) states that information is never bad for the analyst. Indeed, consider an analyst who has already made the identification $V \subseteq \mathcal{U}$—that is, who already knows that the subject's preference is contained in V—and is contemplating the value of an additional observation that would reveal if the subject's preference is in W. The current value of her identification is $\tau(V)$. If she learns the additional observation, the total value will depend on if the subject's preference lies in W or not, resulting in $\tau(W)$ or $\tau(V \setminus W)$ respectively. According to her beliefs, the former occurs with probability $\mu(W \mid V)$ and the latter with probability $1 - \mu(W \mid V)$. Thus, Condition (T1) requires the expected value of this further information is (weakly) positive. Notice that if $\tau(W) \geq \tau(V)$ whenever $W \subseteq V$, then the constraint follows immediately.

In many cases, the analyst may not entertain a prior over $\mathcal{U}$. Nonetheless, our theory applies almost exactly. In this case, the value function F can be written as

$$F(A, \mathscr{P}) = \sum_{P \in \mathscr{P}} \nu(W_{A,P}),$$

for an abstract identification index ν which does not separate the intrinsic value of identification from its likelihood. This is akin to the failure of separation into tastes and beliefs in state-dependent expected utility. In this case, ν must be sub-additive to imbue a positive value for information.

3 Normative Principles of Experimental Design

If $\mathscr{P}$ is a partition of some set X and $Y \subseteq X$, then $\mathscr{P}|_Y = \{P \cap Y \mid P \in \mathscr{P}\}$ is a partition of Y; we denote the corresponding (possibly empty) cells as $P|_Y$. If $\mathscr{P}$ and $\mathscr{Q}$ are both partitions of the same set X and $Y \subseteq X$ is measurable with respect to both $\mathscr{P}$ and $\mathscr{Q}$ then $\mathscr{P}_Y\mathscr{Q}$ denotes the partition that coincides with $\mathscr{P}$ over Y and with $\mathscr{Q}$ over $X \setminus Y$.

We impose four axioms on $\succcurlyeq$, the first of which requires that it admits an expected utility representation. This axiom is not expressed in terms of its individual choices, as its behavioral foundations are widely known.

A1—expected utility. $\succcurlyeq$ entertains an expected utility representation.

The following three axioms reflect our normative principles. Recall that our first principle, Information Monotonicity, asserts that the analyst has a preference for sharper identification. In the current domain, this amounts to assuming finer partitions will always be weakly preferred..

A2—monotonicity. For $A \in \mathbb{D}$, and partitions $\mathscr{P}$, $\mathscr{Q}$ of A it follows that

$$(A, \mathscr{P}) \succcurlyeq (A, \mathscr{Q})$$

whenever $\mathscr{P}$ is finer than $\mathscr{Q}$.

Our second principle maintains that the value of an experiment should only depend on what is potentially identifiable. Thus, it requires indifference between two experiments that have the same ex-ante identifiable set's of preferences (up to μ-probability 0 events).

A3—STRUCTURAL INVARIANCE. Let $(A, \mathscr{P})$ and $(B, \mathscr{Q})$ be such that $\{W_{A,P}|P \in \mathscr{P}\}$ is μ-equivalent to $\{W_{B,Q}|Q \in \mathscr{Q}\}$. Then $(A, \mathscr{P}) \sim (B, \mathscr{Q})$.

Finally, Identification Separability demands that the value of some partial identification cannot depend on the counterfactual. We take advantage of our lottery domain to capture this. Specifically, fix a decision problem A and partitions $\mathscr{P}$ and $\mathscr{Q}$ of A. Identification Separability requires that for any subset $B \subseteq A$, the value of identification given $(A, \mathscr{P})$ and $(A, \mathscr{Q})$, conditional that the choice is in B, should only depend on $\mathscr{P}|_B$ and $\mathscr{Q}|_B$, respectively. Hence, if the agent will choose an element of B, a radomized experiment between $(A, \mathscr{P})$ and $(A, \mathscr{Q})$ should be indifferent to a randomized experiment between $(A, \mathscr{P}_B\mathscr{Q})$ and $(A, \mathscr{Q}_B\mathscr{P})$.

A4—IDENTIFICATION SEPARABILITY. For $A \in \mathbb{D}$, partitions $\mathscr{P}$, $\mathscr{Q}$ of A, and $B \subseteq A$ measurable with respect to both $\mathscr{P}$ and $\mathscr{Q}$

$$\tfrac{1}{2}(A, \mathscr{P}) + \tfrac{1}{2}(A, \mathscr{Q}) \sim \tfrac{1}{2}(A, \mathscr{P}_B\mathscr{Q}) + \tfrac{1}{2}(A, \mathscr{Q}_B\mathscr{P}).$$

Requiring that the value of an object that is uncertain does not depend on the counterfactual is a well known implication of Dynamic Consistency and Consequentialism. As we now elaborate, our axiom is a direct implication of these requirements.

Consider an extension of the analyst's preferences to the case in which she knows choice out of A will be contained in B, denoted $\succcurlyeq_B$, and the case in which she knows the opposite, denoted $\succcurlyeq_{B^c}$. If the choice is in B, then in terms of preference identification, the experiment $(A, \mathscr{P})$ is equivalent to $(A, \mathscr{P}_B\mathscr{Q})$ and $(A, \mathscr{Q})$ to $(A, \mathscr{Q}_B\mathscr{P})$. Analogously, if the choice is not in B, $(A, \mathscr{Q})$ is equivalent to $(A, \mathscr{P}_B\mathscr{Q})$ and $(A, \mathscr{P})$ to $(A, \mathscr{Q}_B\mathscr{P})$. Hence, if the analyst's preference do not depend on the counterfactual

(consequentialism), then

$$(A, \mathscr{P}) \sim_B (A, \mathscr{P}_B \mathscr{Q}) \quad \text{and} \quad (A, \mathscr{Q}) \sim_B (A, \mathscr{Q}_B \mathscr{P});$$
$$(A, \mathscr{Q}) \sim_{B^c} (A, \mathscr{P}_B \mathscr{Q}) \quad \text{and} \quad (A, \mathscr{P}) \sim_{B^c} (A, \mathscr{Q}_B \mathscr{P}).$$

Therefore, under Independence,

$$\tfrac{1}{2}(A, \mathscr{P}) + \tfrac{1}{2}(A, \mathscr{Q}) \sim_B \tfrac{1}{2}(A, \mathscr{P}_B \mathscr{Q}) + \tfrac{1}{2}(A, \mathscr{Q}_B \mathscr{P})$$
$$\tfrac{1}{2}(A, \mathscr{P}) + \tfrac{1}{2}(A, \mathscr{Q}) \sim_{B^c} \tfrac{1}{2}(A, \mathscr{P}_B \mathscr{Q}) + \tfrac{1}{2}(A, \mathscr{Q}_B \mathscr{P}).$$

Finally, observe that if the analyst's ex-ante preference respects her conditional preferences (dynamic consistency), she must exhibit

$$\tfrac{1}{2}(A, \mathscr{P}) + \tfrac{1}{2}(A, \mathscr{Q}) \sim \tfrac{1}{2}(A, \mathscr{P}_B \mathscr{Q}) + \tfrac{1}{2}(A, \mathscr{Q}_B \mathscr{P}).$$

These four axioms—**A1** providing the expected utility structure, and **A2**–**A4** capturing our three normative principles—characterize expected identification value maximization.

THEOREM 1. *The preference $\succcurlyeq$ satisfies* **A1**–**A4** *if and only if it has an expected identification value representation.*

4 IDENTIFYING EXPECTED UTILITY PREFERENCES

The structural invariance axiom, **A3**, states abstractly that the value of an experiment should not depend on structural details. When the choice environment has a specific structure, this principle can be made concrete so as to reflect the particular invariant quantities of the environment at hand. We will now show how structural invariance captures tangible restrictions on the ranking of experiments within the specific environment of linear utility. Here, the analyst is interested in identifying the Von Neumann–Morgenstern utility index of the subject, under the maintained assumption that he is an expected utility maximizer.

This environment is closely related to the setup of random expected utility models à la Gul and Pesendorfer (2006). In particular, the experimenter's prior μ defines a GP random expected utility model. Our conditions on experiments (E1) and (E2) and our richness condition are then direct consequences of the GP assumptions.

Let Δ be a convex and compact subset of a finite dimensional Euclidean space $\mathbb{R}^\ell$ and $\mathcal{U}^\Delta$ denote the set of expected utility (i.e., affine) functions over Δ. So the set of decision problems $\mathbb{D}$ is the set of all finite subsets of Δ. Let Ω^Δ be the smallest algebra on $\mathcal{U}^\Delta$ that contains all identifiable sets: that is contains $W_{A,B}$ for all $A \in \mathbb{D}$ and $B \subseteq A$. Following GP call a $\mu \in \mathbb{P}(\mathcal{U}^\Delta, \Omega^\Delta)$ *regular* if $\mu(u \in \mathcal{U}^\Delta | u(x) = u(y)) = 0$ for all $x, y \in \Delta$.[8]

Theorem 2, below, shows that our richness assumption is not overly strong; within the expected utility model, it is a natural consequence of standard assumptions over the ex-ante distribution on utilities.

THEOREM 2. *If μ is regular, then $\mathbb{E}^\Delta = \{(A, \mathscr{P}) \mid A \subseteq \Delta$ is finite, $\mathscr{P}$ partitions $A\}$ is a rich set of experiments.*

We will now recast structural invariance in a domain specific manner, illuminating the concrete notion of invariance that is relevant to the expected utility model. Specifically, we will show that structural invariance is equivalent to two axioms that specify when two experiments are equivalent and that do not need to directly reference sets of identifiable utilities. To do this, we first need to define a notion of mixing: For $A, B \subseteq \Delta$, and $\alpha \in [0, 1]$, let $\alpha A + (1-\alpha)B = \{\alpha x + (1-\alpha)y \mid x \in A, y \in B\}$ denote the the Minkowski sum. If A and B are decision problems (i.e., are finite), then so is $\alpha A + (1-\alpha)B$.

Observe that under the assumption that u is linear, if x maximizes u over A and y maximizes u over B, then the mixture of x and y will maximize u over the corresponding mixture of the menus. That is:

$$\left.\begin{array}{l} x \in \arg\max_A u(\cdot) \\ y \in \arg\max_B u(\cdot) \end{array}\right\} \quad \text{if and only if} \quad \alpha x + (1-\alpha)y \in \underset{\alpha A + (1-\alpha)B}{\arg\max}\ u(\cdot)$$

[8] GP shows that regular measures are exactly the measures that can be potentially identified from (random) choice data.

for any $\alpha > 0$ and $A, B \in \mathbb{D}$. Now consider the experiment $(A, \{P_1, \dots P_n\})$ and some other decision problem B; by the above logic $u \in W_{A,P_i}$ if and only $u \in W_{\alpha A+(1-\alpha)B,\alpha P_i+(1-\alpha)B}$. Indeed, there must be some $y \in B$ that maximizes u over B, so that $\alpha x + (1-\alpha)y \in \alpha P_i + (1-\alpha)B$ maximizes u over the mixture. Hence, translating an experiment by mixing both the decision problem and the observability partition with some common B does not alter which preference sets can be identified. This particular form of invariance is captured by the following axiom.[9]

A5—TRANSLATION INVARIANCE. For $A, B \in \mathbb{D}$, we have

$$(A, \{P_1, \dots P_n\}) \sim (\alpha A + (1-\alpha)B, \{Q_1, \dots Q_n\}),$$

whenever $Q_i \subseteq (\alpha P_i + (1-\alpha)B)$ for all $i \leq n$.

Recall that **A3** also implies that the value of identification should not depend on zero probability perturbations. The following axiom reflects this implication.

A6—BELIEF CONSISTENCY. Fix $A \in \mathbb{D}$, and let $\{P_1, P_2, \dots P_n\}$ be a partition of A such that $\mu(W_{A,P_1}) = 0$. Then:

$$(A, \{P_1, P_2, \dots P_n\}) \sim (A, \{P_1 \cup P_2, \dots P_n\}).$$

Within the expected utility framework, translation invariance and belief consistency are equivalent to structural invariance. Thus, along with the other axioms from the general model, the two axioms above provide a characterization of expected identification value maximization with linear utility.

THEOREM 3. *Let $\succcurlyeq$ be defined over $\Pi(\mathbb{E}^\Delta)$. Then $\succcurlyeq$ satisfies **A5** and **A6** if and only if it satisfies **A3**.*

[9]The reason there is a subset, rather than set equality, in the axiom is that it is possible that $z \in \alpha A+(1-\alpha)B$ is not a unique mixture of two elements. That is, $z = \alpha x+(1-\alpha)y = \alpha x'+(1-\alpha)y'$ for some $x, x' \in A$ and $y, y' \in B$. For these elements, the cell of the partition in which they reside is not determined, but, it turns out not to matter. See the appendix for the formal argument.

5 Shannon Entropy

Aside from applying to broader settings, out analysis can be used as a building block to provide foundations for specific theories of Bayesian experimental design. In this section we show how our Structural Invariance and Identification Separability can be strengthened to characterize the case in which the identification index τ conforms to the Shannon entropy:

$$\tau(W) = -\log(\mu(W)).$$

Notice that within this special case, the value of identifying a subset of utilities depends only on it's ex-ante probability.

First, Structural Invariance can be strengthened to a Symmetry axiom stating that experiments inducing more evenly distributed probabilities across observations are preferred. When the analyst's value for identification depends only on its prior probability, then experiments in which the probability of each observation is approximately equal ensure that the ex-post identification value is approximately equal as well. Thus, for a cautious analyst, such experiments are desirable as they increase the worst case identification.

A7—symmetry. Fix $A, B \in \mathbb{D}$, and partitions $\{P_1, \dots P_n\}$ and $\{Q_1, \dots Q_n\}$ of A and B, respectively. Then if $|\mu(W_{B,Q_i}) - \frac{1}{n}| \geq |\mu(W_{A,P_i}) - \frac{1}{n}|$ for $i \leq n$, it follows that

$$(A, \{P_1 \dots P_n\}) \succcurlyeq (B, \{Q_1 \dots Q_n\}).$$

Notice that if two experiments induce μ-equivalent identification sets then they also induce the same distribution over the set of positive probability observations. Under **A6**, we can ignore μ-probability zero observations, and so symmetry ensures that the experiments are equally valued. In other words, Symmetry (along with Belief Consistency) imply Structural Invariance.

Next, we can strengthen Identification Separability to get not only additivity but the specific logarithmic form of the entropic representation. Entropic Additivity,

below, disciplines how much the analyst values replacing P_1 in $(A, \mathscr{P})$ with one of it partitions. Let $\mathscr{P} = \{P_1, \dots P_n\}$ be a partition A and $\mathscr{P}_1 = \{P_1^1, \dots P_k^1\}$ a partition P_1. Then $\mathscr{P}^\dagger = \{P_1^1, \dots P_k^1, P_2, \dots P_n\}$ is also partition of A. Fix an experiment $(B, \mathcal{Q})$ such that $\mu(W_{B,Q_i}) = \mu(W_{A,P_i^1} \mid W_{A,P_1})$ for $i = 1, ..., k$.

The fundamental character of the entropic representation is that the value of an experiment only depends on the ratio between the prior and each induced posterior: as such learning which element of $\mathcal{Q}$ was chosen would impart the same value to the analyst as learning which element of $\mathscr{P}_1$ was chosen *conditional on already knowing that* P_1 *was chosen from* $\mathscr{P}$. Further, the partition $\mathscr{P}^\dagger$ is exactly like learning $\mathscr{P}$ and in the event $P_1 \in \mathscr{P}$ is chosen further learning which element of $\mathscr{P}_1$ is chosen. The extra learning happens with probability $\mu(W_{A,P_1})$: so the value $(A, \mathscr{P}^\dagger)$ should equal the value of $(A, \mathscr{P})$ plus $\mu(W_{A,P_1})$ times the value of learning the element chosen from $\mathscr{P}_1$, which as argued above is the value of $(B, \mathcal{Q})$. Translating this into lotteries, we have:

A8—ENTROPIC ADDITIVITY. Fix $A \in \mathbb{D}$ let $\mathscr{P} = \{P_1, \dots P_n\}$ partition A and let $\{P_1^1, \dots P_k^1\}$ partition P_1. So $\mathscr{P}^\dagger = \{P_1^1, \dots P_k^1, P_2, \dots P_n\}$ is also partition of A. Set $\alpha = \frac{1}{1+\mu(W_{A,P_1})}$. Then if $B \in \mathbb{D}$ is such that $\mathcal{Q} = \{Q_1, \dots Q_k\}$ is a partition of B with $\mu(W_{B,Q_i}) = \mu(W_{A,P_i^1} \mid W_{A,P_1})$, it follows that

$$\alpha(A, \mathscr{P}^\dagger) + (1-\alpha)(A, \{A\}) \sim \alpha(A, \mathscr{P}) + (1-\alpha)(B, \mathcal{Q})$$

By replacing Structural Invariance and Identification Separability with the stronger entropic variants above, we find a characterization of expected entropy minimization.

THEOREM 4. *Let* μ *be non-atomic. The preference* $\succcurlyeq$ *satisfies* **A1**–**A2** *and* **A6**–**A8** *if and only if it is represented by*

$$F(\pi) = -\sum_{\text{supp}(\pi)} \Big(\sum_{P \in \mathscr{P}} \log(\mu(W_{A,P}))\mu(W_{A,P}) \Big) \pi(A, \mathscr{P}).$$

While the Shannon specification has significant normative appeal, Symmetry does impose restrictions on the analyst's risk attitudes that need to be spelled out.

To illustrate, consider two experiments $(A, \{P_1, P_2\})$ and $(B, \{Q_1, Q_2\})$. Suppose

$$\mu(W_{A,P_1}) = \mu(W_{A,P_2}) = \frac{1}{2}$$
$$\mu(W_{B,Q_1}) = \frac{3}{4}, \mu(W_{B,Q_2}) = \frac{1}{4}.$$

Thus, if the analyst offers $(A, \{P_1, P_2\})$ she will be able to rule out "half" of the preference for the subject regardless of the subject's true utility. However, if she offers $(B, \{Q_1, Q_2\})$, then the size of the mass of preference she will be able to eliminate depends on the subjects preference. If the subject's preference is maximized in Q_1, she will be able to eliminate three quarters; if it is maximized in Q_2, she will only eliminate one quarter. The entropic model imposes that the former is preferred, implicitly requiring a specific risk preference on the part of the analyst. We view this as beyond the scope of what can be argued only on normative grounds.

6 Belief Free Models

In many cases, the analyst may not entertain a prior over $\mathcal{U}$. Nonetheless, our theory applies almost exactly. In this case, the value function F can be written as

$$F(A, \mathscr{P}) = \sum_{P \in \mathscr{P}} \nu(W_{A,P}), \tag{1}$$

for an abstract subadditive identification index ν which does not separate the intrinsic value of identification from its likelihood. This is akin to the failure of separation into tastes and beliefs in state-dependent expected utility.

In the original model, the value of identification was invariant to μ-measure zero perturbations. This is what allowed us to work with μ-equivalent-approximations of partitions of $\mathcal{U}$, greatly extending the set of scope of application. Without beliefs, we must re-cast the notation of null sets in a more general from. Call $V \in \Omega$ *transparent* if for any $(A, \{P_1, P_2, \ldots P_n\})$ with $W_{A,P_1} = V$, we have

$$(A, \{P_1, P_2, \ldots P_n\}) \sim (A, \{P_1 \cup P_2, \ldots P_n\}). \tag{2}$$

Using transparency as a preference based definition of nullness, we can restate the condition (E2) and richness without having to appeal to beliefs. In particular, assume that for a given set of experiments $\mathbb{E}$

(E1') For all $(A, \mathscr{P}) \in \mathbb{E}$ and $P, Q \in \mathscr{P}$, if $V \subseteq (W_{A,P} \cap W_{A,Q})$, then V is transparent.

Further, call two (finite) collections of subsets of $\mathcal{U}$, $\{W_1, \ldots W_n\}$ and $\{V_1, \ldots W_m\}$ *T-equivalent* if for every non-transparent W_i there exists a V_j such that $W_i \setminus V_j$ and $V_j \setminus W_i$ are both transparent, and vice versa. That is, the collections are T-equivalent if they can be identified up to transparent sets.

Modulo these two changes, Theorem 1 goes through exactly as stated to arrive at a representation of the form (1). To see this, notice that the set of all $\{V \mid V \subseteq W_{A,P} \cap W_{A,Q}, \text{ for some } (A, \mathscr{P}) \in \mathbb{E}, P, Q \in \mathscr{P}\}$ is a down-set. The ideal generated by this down-set is a subset of all transparent sets (it is immediate from their definition that transparent sets are closed under finite unions). Thus, there exists a $\{0, 1\}$-valued finitely additive measure on Ω sending all such sets to 0. We can then define μ as this measure and set $\nu = \mu \cdot \tau$.

7 Observability Constraints

We conclude the paper by illustrating how our framework is general enough to capture partial observability in dynamic environments. We begin by considering the case in which an analyst employs an adaptive method and the subject is an expected utility maximizer.

Suppose an analyst first offers $\{x_0, y_0\}$. If the subject chooses x_0, then she offers $\{x_x, y_x\}$, otherwise she offers $\{x_y, y_y\}$. Given the nature of the procedure, if the agent chooses x (y) from $\{x, y\}$, then the analyst will know the choice out of $\{x_x, y_x\}$ but not the choice out of $\{x_y, y_y\}$ ($\{x_x, y_x\}$). Figure 3 illustrates the procedure.

Observe that because of linearity of the preferences, observing a choice from

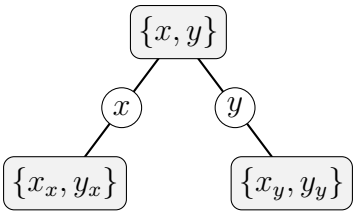

Figure 3: Adaptive Procedure

$\{x, y\}$ and $\{x_x, y_x\}$ is the same as observing a choice from

$$A_x = \{\frac{1}{2}x + \frac{1}{2}x_x, \frac{1}{2}x + \frac{1}{2}y_x, \frac{1}{2}y + \frac{1}{2}x_x, \frac{1}{2}y + \frac{1}{2}y_x\}.$$

The reason is that any expected utility maximize that would choose a out of $\{x, y\}$ and b out of $\{x_x, y_x\}$ would choose $\frac{1}{2}a + \frac{1}{2}b$ out of A_x.

Similarly, observing a choice from $\{x, y\}$ and $\{x_y, y_y\}$ is the same as observing a choice from

$$A_y = \{\frac{1}{2}x + \frac{1}{2}x_x, \frac{1}{2}x + \frac{1}{2}y_y, \frac{1}{2}y + \frac{1}{2}x_y, \frac{1}{2}y + \frac{1}{2}y_y\}.$$

Consider the menu $A_x \cup A_y$ and the partition

$$\mathscr{P} = \Big\{\{\frac{1}{2}x + \frac{1}{2}x_x, \frac{1}{2}x + \frac{1}{2}y_x\}, \{\frac{1}{2}y + \frac{1}{2}x_y, \frac{1}{2}y + \frac{1}{2}y_y\}\Big\}.$$

Then the information provided by the adaptive design is equivalent to the information provided by $(A_x \cup A_y, \mathscr{P})$.

The above example can be easily generalized to adaptive procedures that employ T rounds of choices that allow for non-binary menus. While the intuition is clear, precisely stating an equivalence result requires a significant amount of notation and so we leave this at the informal level.

Next, consider an analyst interested in learning the subject's preference by employing a dynamic game. Suppose the game features two players, the subject and a computer. The subject can first choose *out* (o) or *in* (i). If the subject chooses in, then the computer randomizing between *right* (r) and *left* (l). Following left the subject has a choice between a and b and following right, a choice between c and d.

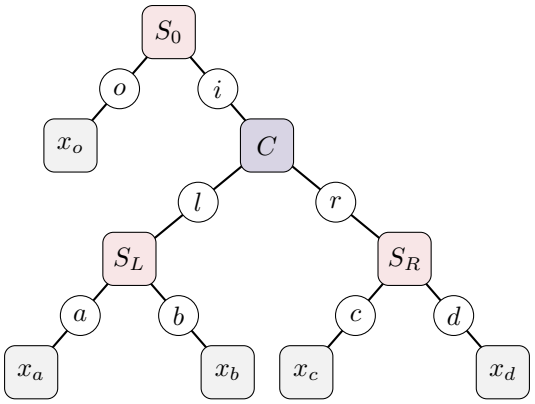

Figure 4: The dynamic game between the subject and computer. The subject's decision nodes are shaded red and the computer's blue.

The game is provided in Figure 4.

Suppose that the analyst only observes the on-path strategies; she cannot know how the subject would behave in a sub-game that is not reached. There are five actions the analyst could potentially observe: (o), (i, a), (i, b), (i, c), and (i, d). Notice which of these is observed depends not only on the subject's choice but also the outcome of the computer randomization. These observations can be adapted into our framework. Let A be the set of all strategies for the subject:

$$A = \left\{ \begin{array}{l} (i,a,c),(i,b,c),(i,a,d),(i,b,d), \\ (o,a,c),(o,a,d),(o,b,c),(o,b,d) \end{array} \right\}$$

Now consider the following partitions of A

$$\mathscr{P}_L = \left\{ \begin{array}{l} \left\{ \begin{array}{l} (i,a,c),(i,b,c), \\ (i,a,d),(i,b,d) \end{array} \right\}, \\ \left\{ (o,a,c),(o,a,d) \right\}, \\ \left\{ (o,b,c),(o,b,d) \right\} \end{array} \right\} \quad \mathscr{P}_R = \left\{ \begin{array}{l} \left\{ \begin{array}{l} (i,a,c),(i,b,c), \\ (i,a,d),(i,b,d) \end{array} \right\}, \\ \left\{ (o,a,c),(o,b,c) \right\}, \\ \left\{ (o,a,d),(o,b,d) \right\} \end{array} \right\}$$

Observe that a single choice of $(A, \mathscr{P}_L)$ would yield the same information as observing the subject play the dynamic game in the event that the computer chooses

left, and likewise $(A, \mathscr{P}_R)$ should the computer choose *right*. Thus, if the computer chooses *left* with probability α, then an observation in the dynamic game is observationally equivalent to an observation from the randomized experiment $\alpha(A, \mathscr{P}_L) + (1-\alpha)(A, \mathscr{P}_R)$. Again, this can be generalized: observations of on-path behavior in dynamic games can be incorporated into our framework via the appropriately constructed random experiments.

A Proofs

A.1 Proof of Theorem 1

Let $\texttt{part}(\mathcal{U})$ denote the finite Ω-measurable partitions of $\mathcal{U}$. That is, $\{W_1, \dots W_n\} \in \texttt{part}(\mathcal{U})$ if it is a (finite) partition of $\mathcal{U}$ such that each $W_i \in \Omega$.

Lemma 1. *Let $(A, \{P_1, \dots P_n\}) \in \mathbb{E}$; then $\{W_{A,P_i}\}_{i \leq n}$ is μ-equivalent to some partition $\mathcal{W} \in \texttt{part}(\mathcal{U})$.*

Proof. By (E1) $\{W_{A,P_i}\}_{i \leq n}$ and since each $u \in \mathcal{U}$ find some maximum on A, we have $\bigcup_{i \leq n} W_{A,P_i} = \mathcal{U}$. Define $W_i = W_{A,P_i} \setminus \bigcup_{i<j} W_{A,P_j}$. Clearly, $\{W_i\}_{i \leq n} \in \texttt{part}(\mathcal{U})$. Moreover, $\mu(W_{A,P_i}) = \mu(W_{A,P_i}) - \sum_{i<j} \mu(W_{A,P_i} \cap W_{A,P_j}) \leq \mu(W_{A,P_i} \setminus \bigcup_{i<j} W_{A,P_j}) = \mu(W_i) = \mu(W_i \cap W_{A,P_i}) \leq \mu(W_{A,P_i})$, establishing μ-equivalence (the first, and only non-set-theoretically obvious, equality comes form (E2)). ★

Lemma 2. *Let $\boldsymbol{W} \subseteq \Omega$ be μ-equivalent to $\boldsymbol{V} \subseteq \Omega$, and assume $\mu(W \cap W') = 0$ and $\mu(V_i \cap V_j) = 0$ for any distinct $W, W' \in \boldsymbol{W}$ and $V, V' \in \boldsymbol{V}$. Then there exists a bijection, q, between $\{W \in \boldsymbol{W} \mid \mu(W) > 0\}$ and $\{V \in \boldsymbol{V} \mid \mu(V) > 0\}$ such that $\mu(W) = \mu(W \cap h(W)) = \mu(q(W))$.*

Proof. Take some $W \in \boldsymbol{W}$ with $\mu(W) > 0$. By μ-equivalence, there exists a $V \in \boldsymbol{V}$ such that $\mu(W) = \mu(W \cap V) = \mu(V)$. To see that it is unique, let $V, V' \in \boldsymbol{V}$ both be such that the needed relation holds. Then we have $\mu(W) < 2\mu(W) = \mu(V \cap W) + \mu(V' \cap W) = \mu((V \cup V') \cap W) + \mu((V \cap V') \cap W) \leq \mu(W) + \mu(V \cap V')$. This means that $\mu(V \cap V') > 0$. By the condition in the statement of the Lemma, this requires $V = V'$. ★

LEMMA 3. *Let $\mathcal{W} \in \texttt{part}(\mathcal{U})$ and let $(A, \mathscr{P}) \in \mathbb{E}$ be such that such that $\{W_{A,P} | P \in \mathscr{P}\}$ is μ-equivalent it. Then there exists a family of partitions*

$$\{\mathscr{P}_{\mathcal{V}} \mid \mathcal{V} \text{ is a coarsening of } \mathcal{W}\}$$

such that

1. *$\{W_{A,P} | P \in \mathscr{P}_{\mathcal{V}}\}$ is μ-equivalent $\mathcal{V}$, and*
2. *If $\mathcal{V}'$ is a coarsening of $\mathcal{V}$, then $\mathscr{P}_{\mathcal{V}'}$ is a coarsening of $\mathscr{P}_{\mathcal{V}}$.*

Proof. First, for each $W \in \mathcal{W}$ with $\mu(W) > 0$, let $P_W \in \mathscr{P}$ be the unique element such that $\mu(W_{A,P}) = \mu(W_{A,P} \cap W) = \mu(W)$. This exists by Lemma 2.

Now, for each $V \in \mathcal{V}$, let $[V] = \{P_W \in \mathscr{P} \mid W \in \mathcal{W}, \mu(W) \geq 0, W \subseteq V\}$. It is easy to see that $\{\bigcup_{P_W \in [V]} W_{A,P_W} \mid V \in \mathcal{V}\}$ is μ-equivalent $\mathcal{V}$. Indeed, either $\mu(V) = 0$, in which case $[V] = \varnothing$ or $\mu(V) = \sum_{W \subseteq V} \mu(W) = \sum_{P_W \in [V]} \mu(W_{A,P_W}) = \mu(\bigcup_{P_W \in [V]} W_{A,P_W})$. Let $P_{\mathcal{V}}$ be the coarsest partition containing $\{\bigcup [V] \mid V \in \mathcal{V}\}$. Note that $\bigcup [V] \cap \bigcup [V'] = \varnothing$ whenever $V \neq V'$, so $P_{\mathcal{V}}$ will simply be $\{\bigcup [V] \mid V \in \mathcal{V}\}$ adjoined with whatever elements of A were not in any $\bigcup [V]$—these are exactly the observations that have 0-probability. ★

Proof of Theorem 1. From **A1** there exists a vNM index $\texttt{vnm} : \mathbb{E} \to \mathbb{R}$ such that

$$\pi \succcurlyeq \rho \qquad \iff \qquad \sum_{\text{supp}(\pi)} \texttt{vnm}(e)\pi(e) \geq \sum_{\text{supp}(\rho)} \texttt{vnm}(e)\rho(e)$$

$\texttt{vnm}$ can be chosen such that $\texttt{vnm}(\{x\}, \{\{x\}\}) = 0$ (since this induces the trivial partition of $\mathcal{U}$ independent of x, by **A3**, the choice of x is irrelevant).

For each $\mathcal{W} \in \texttt{part}(\mathcal{U})$ let $(A_{\mathcal{W}}, \mathscr{P}_{\mathcal{W}}) \in \mathbb{E}$ be such that such that $\{W_{A_{\mathcal{W}},P} | P \in \mathscr{P}_{\mathcal{W}}\}$ is μ-equivalent to $\mathcal{W}$. This exists by the richness assumption on the set of experiments. Define the function $\varphi : \texttt{part}(\mathcal{U}) \to \mathbb{R}$ as

$$\varphi : \mathcal{W} \mapsto \texttt{vnm}(A_{\mathcal{W}}, \mathscr{P}_{\mathcal{W}}). \tag{3}$$

By Axiom 3, φ does not depend on the choice of $(A_{\mathcal{W}}, \mathscr{P}_{\mathcal{W}})$.

Call $\nu : \Omega \to \mathbb{R}$ a *GL-representation* of φ if $\nu(\varnothing) = 0$ and

$$\varphi(\mathcal{W}) = \sum_{W \in \mathcal{W}} \nu(W). \tag{4}$$

LEMMA 4. *A GL representation of φ exists.*

Proof. Following Gilboa and Lehrer (1991) (Observation 2.1), call two partitions, $\mathcal{W}, \mathcal{V} \in \texttt{part}(\mathcal{U})$, *non-intersecting* iff there is an event $U \in \Omega$ such that U is measurable with respect to both $\mathcal{W}$ and $\mathcal{V}$ and such that $\mathcal{W}|_U$ refines $\mathcal{V}|_U$ and $\mathcal{V}|_{U^c}$ refines $\mathcal{U}|_{U^c}$.

Theorem 3.2 of Gilboa and Lehrer (1991) states that a GL-representation of φ exists if and only if

$$\varphi(\mathcal{W} \wedge \mathcal{V}) + \varphi(\mathcal{W} \vee \mathcal{V}) = \varphi(\mathcal{W}) + \varphi(\mathcal{V}) \tag{5}$$

for any non-intersecting partitions, where $\mathcal{W} \wedge \mathcal{V}$ and $\mathcal{W} \vee \mathcal{V}$ denote their meet (coarsest common refinement) and join (finest common coarsening), respectively.

So, let $\mathcal{W}$ and $\mathcal{V}$ be non-intersecting and U the jointly measurable event delineating which partition is finer. Consider an experiment $(A, \mathscr{P}_{\mathcal{W}\wedge\mathcal{V}})$ such that $\{W_{A,P} | P \in \mathscr{P}_{\mathcal{W}\wedge\mathcal{V}}\}$ is μ-equivalent to $\mathcal{W}$. Again, this exists by the richness assumption.

By Lemma 3,we can construct partitions $\mathscr{P}_{\mathcal{W}}$, $\mathscr{P}_{\mathcal{V}}$, $\mathscr{P}_{\mathcal{W}\vee\mathcal{V}}$, $\mathscr{P}_{\{U,U^c\}}$ of A, each inducing a partition μ-equivalent with respect to the corresponding partition of $\mathcal{U}$ (i.e., indicated by the subscript).

Let $B \in \mathscr{P}_{\{U,U^c\}}$ be the cell such that $\mu(W_{A,B}) = \mu(W_{A,B} \cap U) = \mu(U)$. It follows that B is measurable with respect to all of the above partitions, and furthermore, $(\mathscr{P}_{\mathcal{W}\wedge\mathcal{V}})_B(\mathscr{P}_{\mathcal{W}\vee\mathcal{V}}) = \mathscr{P}_{\mathcal{W}}$ and $(\mathscr{P}_{\mathcal{W}\vee\mathcal{V}})_B(\mathscr{P}_{\mathcal{W}\wedge\mathcal{V}}) = \mathscr{P}_{\mathcal{V}}$.

Applying axiom **A4**, we have

$$\tfrac{1}{2}(A, \mathscr{P}_{\mathcal{W}\wedge\mathcal{V}}) + \tfrac{1}{2}(A, \mathscr{P}_{\mathcal{W}\vee\mathcal{V}}) \sim \tfrac{1}{2}(A, \mathscr{P}_{\mathcal{W}}) + \tfrac{1}{2}(A, \mathscr{P}_{\mathcal{V}})$$

Thus, from (3), the definition of φ, we obtain (5). Theorem 3.2 of Gilboa and Lehrer (1991) ensures us of the existence of some GL-representation. ★

Call $V \in \Omega$ *transparent* if for any partition containing V, $\{V, W, U_1 \dots U_n\}$, it follows that

$$\varphi(\{V, W, U_1 \dots U_n\}) = \varphi(\{V \cup W, U_1 \dots U_n\}). \tag{6}$$

Notice, while it is easier to write this as a condition on φ, it is completely determined by the preference. Let $\Omega^{\varnothing}$ collect the transparent measurable sets. Notice that if ν' is any GL-representation of φ and $V \in \Omega^{\varnothing}$ and $W \in \Omega$ with $W \cap V = \varnothing$, it holds that

$$\nu'(W \cup V) = \nu'(W) + \nu'(V). \tag{7}$$

This follows immediately from plugging the GL-representation, (4), into (6). In particular, notice that ν' is finitely additive over $\Omega^{\varnothing}$.

LEMMA 5. *Let $\mathscr{C} \subseteq \varnothing$ be a set of transparent subsets such that (i) $\varnothing \in \mathscr{C}$, (ii) $\mathcal{U} \notin \mathscr{C}$, (iii) $W, V \in \mathscr{C}$ implies $W \cup V \in \mathscr{C}$, and, (iv) $W \in \mathscr{C}$ and $V \in \Omega$ with $V \subseteq W$ implies $V \in \mathscr{C}$. Then there exists a GL-representation, ν, of φ such that $\nu(V) = 0$ for all $V \in \mathscr{C}$.*

Proof. Let $\nu' : \Omega \to \mathbb{R}$ be a GL-representation, which exists by Lemma 4. Notice that $\mathscr{C}$ is a ring of sets, since if $W, V \in \mathscr{C}$ then $W \setminus V \subseteq W \in \mathscr{C}$ by (iv). Further, by (7), it follows that $\nu'|_{\mathscr{C}} : \mathscr{C} \to \mathbb{R}$ is finitely additive. Hence, by Theorem 3.2.5 of Rao and Rao (1983), there exists a finitely additive measure $\mu' : \Omega \to \mathbb{R}$ that extends $\nu'|_{\mathscr{C}}$.

Notice also that $\mathscr{C}$ is an ideal in Ω (as a Boolean algebra of sets). Thus by the Boolean prime ideal theorem, $\mathscr{C}$ is contained in some maximal (proper) ideal, $\mathcal{I} \subset \Omega$. Then

$$\mu'' : W \mapsto \begin{cases} 0 & \text{if } W \in \mathcal{I} \\ \mu'(\mathcal{U}) & \text{otherwise} \end{cases}$$

is a finitely additive measure. It follows that $\mu^\dagger = \mu' - \mu''$ is a finitely additive measure and $\mu^\dagger(\mathcal{U}) = 0$. By Proposition 3.3 of Gilboa and Lehrer (1991), $\nu = \nu' - \mu^\dagger$ is also a GL-representation of φ. Moreover, for $V \in \mathscr{C}$, we have $\nu(V) = \nu'(V) - \mu^\dagger(V) = \nu'(V) - \mu'(V) + \mu''(V) = \nu'(V) - \nu'(V) + 0 = 0$. ★

Let $\mathscr{C}^{null} = \{W \in \Omega \mid \mu(W) = 0\}$. Note that $\mathscr{C}^{null}$ satisfies the conditions for Lemma 5 (that $\mathscr{C}^{null} \subseteq \Omega^{\varnothing}$ is a straightforward consequence of **A3**). Thus, we can

choose ν such that $\mu(W) = 0$ implies $\nu(W) = 0$. Now define τ as

$$\tau : W \mapsto \begin{cases} \frac{\nu(W)}{\mu(W)} & \text{if } W \notin \mathscr{C}^{null} \\ 0 & \text{otherwise.} \end{cases}$$

It is obvious that this τ satisfies (T2). To see that it also τ satisfies (T1), first note that by **A2**, if $\mathcal{W}$ refines $\mathcal{V}$ then $\varphi(\mathcal{W}) \geq \varphi(\mathcal{V})$, and so by Observation 4.1 of Gilboa and Lehrer (1991) ν is subadditive.

Now let $\mu(V) > 0$ and $W \subset V$; by sub-additivity $\nu(W) + \nu(V \setminus W) \geq \nu(V)$. Plugging in for the definition of τ we have $\tau(W)\mu(W) + \tau(V \setminus W)\mu(V \setminus W) \geq \tau(V)\mu(V)$. Dividing by $\mu(V)$ delivers the inequality part of (T1). If we further assume that $\mu(W) = 0$, then $W \in \Omega^{\varnothing}$ and the equality part of (T1) follows from the definition of transparency (7).

Finally, to see that τ represents $\succcurlyeq$ according to ($\star\star$), let $(A, \mathcal{Q}) \in \mathbb{E}$. By Lemma 1, there is some $\mathcal{W} \in \texttt{part}(\mathcal{U})$ such that $\{W_{A,Q}\}_{Q \in \mathcal{Q}}$ is μ-equivalent to $\mathcal{W}$. Let $(A_{\mathcal{W}}, \mathscr{P}_{\mathcal{W}})$ be the experiment used to define $\varphi(\Omega)$. Clearly, $(A, \mathcal{Q})$ and $(A_{\mathcal{W}}, \mathscr{P}_{\mathcal{W}})$ are themselves μ-equivalent, and hence by **A3**, $(A, \mathcal{Q}) \sim (A_{\mathcal{W}}, \mathscr{P}_{\mathcal{W}})$.

By μ-equivalence, for each $Q \in \mathcal{Q}$ with $\mu(W_{A,Q}) > 0$ there exists some $W^Q \in \mathcal{W}$ such that $\mu(W_{A,Q}) = \mu(W^Q \cap W_{A,Q}) = \mu(W^Q)$. In particular, this implies that $\mu(W^Q \setminus W_{A,Q}) = \mu(W_{A,Q} \setminus W^Q) = 0$. This further implies, via the equality part of (T1), that $\tau(W^Q) = \tau(W^Q \cup W_{A,Q}) = \tau(W_{A,Q})$. So finally, we have

$$\begin{aligned} \sum_{Q \in \mathcal{Q}} \tau(W_{A,Q})\mu(W_{A,Q}) &= \sum_{\substack{Q \in \mathcal{Q} \\ \mu(W_{A,Q}) > 0}} \tau(W^Q)\mu(W^Q) \\ &= \sum_{W \in \mathcal{W}} \nu(W) \\ &= \varphi(\mathcal{W}) \\ &= \texttt{vnm}(A_{\mathcal{W}}, P_{\mathcal{W}}) \\ &= \texttt{vnm}(A, Q) \end{aligned}$$

as desired. ■

A.2 Proof of Theorem 2

Preliminaries: Convex Spaces

For a set $X \subseteq \mathbb{R}^\ell$, let $\text{conv}(X)$, $\text{int}(X)$, and $\text{cl}(X)$ denote the convex hull, the interior, and the closure of X, respectively. If X is convex, then $\text{ext}(X)$ collects all the extreme points of X and $\text{ri}(X)$ denotes the relative interior of X. When it is not confusing to do so, we will write $\text{ri}(X)$ and $\text{ext}(X)$ to mean $\text{ri}(\text{conv}(X))$ and $\text{ext}(\text{conv}(X))$ for non-convex X.

For convex X, let $F \subset X$ be called a face if whenever $\alpha x + (1-\alpha)y \in F$ (for $x, y \in X$) then also $x, y \in F$. Let $\mathbb{F}(X)$ denote the set of all (non-empty) faces of X and $\mathbb{F}^\circ(X) = \{\text{ri}(F) \mid F \in \mathbb{F}(X)\}$.

If $X \subseteq \mathbb{R}^\ell$ is a convex set and $\text{ext}(X)$ is finite then X is a called at polytope. Let $\texttt{poly}$ denote the set of all polytopes in $\mathbb{R}^\ell$. If $X \in \texttt{poly}$, then $\mathbb{F}(X)$ is finite.

If $K \subseteq \mathbb{R}^\ell$ and $\lambda K \subseteq K$ for all $\lambda \geq 0$ then K is called a *cone*. We say a cone K is generated by X if $K = \{\lambda x \mid x \in X, \lambda \geq 0\}$. A cone K is *polyhedral* if it is generated by a polytope; let $\mathcal{K}^*$ denote all such cones. Let $\mathcal{K}$ denote the set of pointed polyhedral cones, those cones with $\mathbf{0} \in \text{ext}(K)$. The face of a polyhedral cone is a polyhedral cone.

For $X \in \texttt{poly}$, let

$$X^\star = \bigcup_{I \subseteq \text{ext}(X)} \sum_{i \in I} \frac{x_i}{|I|}$$

The set $X^\star$ is a decision problem that contains one point in the relative interior of every face of X. Further, given a partition of $\mathscr{H} = \{H_1 \dots H_n\}$ of $\mathbb{F}(X)$, let $\mathscr{H}^\star = \{H_1^\star \dots H_n^\star\}$ denote the partition of $X^\star$ defined via $H^\star = X^\star \cap (\bigcup_{F \in H_i} \text{ri}(F))$.

For $X \subseteq \mathbb{R}^\ell$ (convex or not) and $x \in X$ let $N(X,x) = \{u \in \mathcal{U} \mid u(y-x) \leq 0, \text{ for all } y \in X\}$ denote the normal cone of X at x. Alternatively, $N(X,x) = \{u \in \mathcal{U}^\Delta \mid x \in \arg\max_X u\}$. Notice that $W_{A,B} = \bigcup_{x \in B} N(A,x)$.

For $X \in \texttt{poly}$, and a face $F \in \mathbb{F}(X)$, let $N(X,F) = \bigcap_{x \in F} N(X,x)$. It follows that $N(X,F) = \{u \in \mathcal{U}^\Delta \mid F \subseteq \arg\max_{x \in D} u(x)\} = N(X,x)$ for any $x \in \text{ri}(F)$. Notice therefore that

$$\bigcup_{x \in H_i^\star} \text{ri}(N(X^\star, x)) = \bigcup_{F \in H_i} \text{ri}(N(X,F)) \tag{8}$$

It is immediate that $N(X, F)$ is closed and in $\mathcal{K}^*$. Let $\mathcal{N}(X) = \{N(X, F) \mid F \in \mathbb{F}(X)\}$ denote the normal fan of X; that $\mathcal{N}(X)$ is a fan indicates that it is a family of cones such, with the following two properties:

(i) Every nonempty face of a cone in $\mathcal{N}(X)$ is also a cone in $\mathcal{N}(X)$.
(ii) The intersection of any two cones in $\mathcal{N}(X)$ is a face of both.

Furthermore, $\mathcal{N}(X)$ is complete (the union of $\mathcal{N}(X)$ is $\mathcal{U}^\Delta$). Let $\mathcal{N}^\circ(X) = \{\text{ri}(N(X, F)) \mid F \in \mathbb{F}(X)\}$.

LEMMA 6. *The following are true for all convex X and Y:*

1. $\text{cl}(\text{ri}(X)) = \text{cl}(X)$ *(Theorem 6.3 of Rockafellar (1970)).*
2. $\mathbb{F}^\circ(X)$ *is a partition of X (Theorem 18.2 of Rockafellar (1970)).*
3. *Let $F \in \mathbb{F}(X)$ and $Y \subseteq X$ be such that $\text{ri}(Y) \cap F \neq \varnothing$, then $Y \subseteq F$. (Theorem 18.1 of Rockafellar (1970)).*

LEMMA 7. *For $X \in \texttt{poly}$, $\mathcal{N}^\circ(X)$ is a partition of $\mathcal{U}^\Delta$.*

Proof. Let $u \in \mathcal{U}^\Delta$. Since $\mathcal{N}(X)$ is complete, $u \in K$ for some $K \in \mathcal{N}(X)$. By property (i) of fans, we see that $\mathbb{F}(K) \subseteq \mathcal{N}(X)$; since $\mathbb{F}^\circ(K)$ is a partition of K, it follows that $x \in \text{ri}(F)$ for some $F \in \mathcal{N}(X)$. So, the elements of $\mathcal{N}^\circ(X)$ cover $\mathcal{U}^\Delta$.

Now assume that $x \in \text{ri}(K) \cap \text{ri}(K')$ for some $K, K' \in \mathcal{N}(X)$. Then by property (ii) of fans, $K \cap K' \in \mathbb{F}(K)$; moreover, since $x \in (K \cap K') \cap \text{ri}(K) \neq \varnothing$, Lemma 6.3 delivers that the face $K \cap K'$ must be equal to K itself. By symmetry also $K' = (K \cap K') = K$. So, the elements of $\mathcal{N}^\circ(X)$ are disjoint. ■

LEMMA 8. *For polytopes X and X', the following are equivalent*

(i) $X = \alpha X' + Z$, *for some polytope Z and $\alpha > 0$*
(ii) for all $K \in \mathcal{N}(X)$ there is a $K' \in \mathcal{N}(X')$ such that $K \subseteq K'$
(iii) $\mathcal{N}^\circ(X)$ *refines* $\mathcal{N}^\circ(X')$.

Proof. (i) $\leftrightarrow$ (ii) Theorem 15.1.2 of Grnbaum (2003)

(ii) $\rightarrow$ (iii) Take some $\text{ri}(K) \in \mathcal{N}^\circ(X)$. Let $K^\dagger = \bigcap\{K' \in \mathcal{N}(X') \mid K \subseteq K'\}$, which is an element of $\mathcal{N}(X')$ by (ii) and the properties of fans. Moreover, by

construction $K \not\subseteq F$ for any $F \in \mathbb{F}(K^\dagger)$ with $F \subsetneq K^\dagger$. Thus, by (the contrapositive of) Lemma 6.3, we have that $\text{ri}(K) \cap \text{ri}(F) \neq \varnothing$ for all such F. Then, since $\mathbb{F}^\circ(K^\dagger)$ partitions $K^\dagger$, it follows that $\text{ri}(K) \subseteq \text{ri}(K^\dagger)$.

(iii) $\rightarrow$ (ii) Take some $K \in \mathcal{N}(X)$. Then by (iii) $\text{ri}(K) \subseteq \text{ri}(K')$ for some $K' \in \mathcal{N}(X')$. Since K and K' are closed, we have $K = \text{cl}(K)$ and $K' = \text{cl}(K')$. Thus, $K = \text{cl}(\text{ri}(K)) \subseteq \text{cl}(\text{ri}(K')) = K'$, where both equalities come from Lemma 6.1 and the inclusion relation from the fact that taking closures is subset preserving. ■

Proof of Theorem 2. The coarsening property is obvious. We will show that any partition can be captured up to μ-equivalence. Let $\mathcal{W} = \{W_1, \ldots W_n\} \in \texttt{part}(\mathcal{U})$. First, from Gul and Pesendorfer (2006) Proposition 6(ii), we can write each $W_i \in \Omega$ as the finite union of elements in $\mathcal{K}$: $W_i = \bigcup_{j=1}^{m_i} \text{ri}(K_i^j)$. Moreover, by Gul and Pesendorfer (2006) Proposition 4, each $K_i^j = N(X_i^j, x_i^j)$ for some polytope X_i^j and $x_i^j \in X_i^j$. Thus $\text{ri}(K_i^j) \in \mathcal{N}^\circ(X_i^j)$.

Let $a = m_1 + \ldots + m_n$ and consider the polytope $X = \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{1}{a} X_i^j$. By Lemma 8, $\mathcal{N}^\circ(X)$ refines each $\mathcal{N}^\circ(X_i^j)$. Let $\mathscr{H} = \{H_1, \ldots, H_n\}$ be a partition of $\mathbb{F}(X)$ defined by

$$H_i = \{F \in \mathbb{F}(X) \mid \text{ri}(N(X, F)) \subseteq W_i\} \tag{9}$$

Now take some $i \leq n$ and $u \in W_i$. So, there exists some $j \leq m_i$ such that $u \in \text{ri}(K_i^j) \in \mathcal{N}^\circ(X_i^j)$. Since $\mathcal{N}^\circ(X)$ is a partition of $\mathcal{U}$, there exists some $F \in \mathbb{F}(X)$ with $u \in \text{ri}(N(X, F))$, and furthermore, since this partition refines $\mathcal{N}^\circ(X_i^j)$, $\text{ri}(N(X, F)) \subseteq \text{ri}(K_i^j) \subseteq W_i$. Hence $F \in H_i$ and so $u \in \bigcup_{F \in H_i} \text{ri}(N(X, F))$. We have established that $W_i \subseteq \bigcup_{F \in H_i} \text{ri}(N(X, F))$, and since the other inclusion is obvious, that $W_i = \bigcup_{F \in H_i} \text{ri}(N(X, F))$. Now, on the basis of (8), we have

$$W_i = \bigcup_{x \in H_i^\star} \text{ri}(N(X^\star, x)) \tag{10}$$

Finally, by Lemma 2 of Gul and Pesendorfer (2006), we know that for μ which satisfies (E2), it must be that $\mu(\text{ri}(N(X_{\mathcal{W}}^\star, H^\star)) = \mu(N(X_{\mathcal{W}}^\star, H^\star)$. Thus, $\{W_{X^\star, H^\star}\}_{H^\star \in \mathscr{H}^\star}$ is μ-equivalent to $\mathcal{W}$. ■

A.3 Proof of Theorem 3 and 4

Proof of Theorem 3. Let $(A, \{P_1, \dots P_n\})$ and $(B, \{Q_1, \dots Q_m\})$ be such that $\{W_{A,P_i}\}_{i \leq n}$ is μ-equivalent to $\{W_{B,Q_i}\}_{i \leq m}$. Furthermore, from Lemma 2, we can assume there are $1 \leq k \leq n$ elements of each partition with positive μ-probability and for each $i \leq k$, $\mu(W_{A,P_i}) = \mu(W_{A,P_i} \cap W_{B,Q_i}) = \mu(W_{B,Q_i})$.

Consider the problem $C = \frac{1}{2}A + \frac{1}{2}B$. For each $i \leq n$, define $R_i \subseteq C$ as $R_i = \{\frac{1}{2}P_i + \frac{1}{2}B\} \cap \text{ext}(C)$. Clearly, we have for each $i \leq n$, $R_i \subseteq \frac{1}{2}P_i + \frac{1}{2}B$; it follows from **A5** that $(A, \{P_1, \dots P_n\}) \sim (C, \{R_1, \dots R_n\})$.

Now for each $i \leq k$, let $R'_i = R_i \cap (\frac{1}{2}P_i + \frac{1}{2}Q_i) \cap \text{ext}(C) = (\frac{1}{2}P_i + \frac{1}{2}Q_i) \cap \text{ext}(C)$. The final equality arises from the fact that each extreme point of C has a unique decomposition as elements of A and B (so that any $x \in (\frac{1}{2}P_i + \frac{1}{2}Q_i) \cap \text{ext}(C)$ was not in R_j for $j < i$). We claim that $\mu(W_{C,R_i \setminus R'_i}) = 0$. Indeed,

$$\begin{aligned} W_{C,R_i \setminus R'_i} &\subseteq W_{A,P_i} \cap \bigcup_{j \neq i} W_{B,Q_j} \\ &= \bigcup_{j \neq i} (W_{A,P_i} \cap W_{B,Q_j}) \\ &\subseteq \bigcup_{j \neq i} \left((W_{A,Q_i} \cap W_{B,Q_j}) \cup (W_{A,P_i} \setminus W_{B,Q_i}) \right) \end{aligned}$$

The claim then follows from the fact that for all $i \neq j$, $\mu(W_{A,Q_i} \cap W_{B,Q_j}) = 0$ (from (E2)) and $\mu(W_{A,P_i} \setminus W_{B,Q_i}) = 0$ (from μ-equivalence).

By repeatedly appealing to **A6**, we can see that

$$(C, \{R_1, \dots R_n\}) \sim (C, \{R'_1, \dots R'_k, R^\dagger\}),$$

where $R^\dagger = C \setminus \bigcup_{i \leq k} R'_i$. We make use the fact for $i > k$, $\mu(W_{C,R_i}) = 0$ on account of the fact that $W_{C,R_i} \subseteq W_{A,P_i}$. Thus we have

$$(A, \{P_1, \dots P_n\}) \sim (C, \{R_1, \dots R_n\}) \sim (C, \{R'_1, \dots R'_k, R^\dagger\})$$

A symmetric argument ensures that also $(B, \{Q_1, \dots Q_m\}) \sim C, \{R'_1, \dots R'_k, R^\dagger\})$, and so the two experiments are indifferent, as is required. ∎

Proof of Theorem 4. As before let $\texttt{vnm}$ be the utility index that represents $\succcurlyeq$, that exists by **A1**, normalized such that $\texttt{vnm}(\{x\},\{\{x\}\}) = 0$; by **A7**, the choice of x is irrelevant, and by **A2** the trivial experiment is the worst possible, so $\texttt{vnm} : \mathbb{E} \to \mathbb{R}_+$ takes only weakly positive values.

Let $\texttt{prob}$ denote the set of finitely valued probability distributions, i.e., finite lists taking values in $[0,1]$ whose entries sum to 1. Let $\texttt{prob}^* \subset \texttt{prob}$ denote those whose entries are all strictly positive. Define $\zeta : \mathbb{E} \to \texttt{prob}$ as $\zeta(A,\{P_1,\dots P_n\}) = (\mu(W_{A,P_1}),\dots,\mu(W_{A,P_n}))$.

LEMMA 9. *For each* $\{p_1,\dots p_n\} \in \texttt{prob}^*$, *there exists some* $(A,\mathscr{P}) \in \mathbb{E}$ *such that* $\zeta(A,\mathscr{P}) = \{p_1,\dots p_n\}$. *Moreover, if* $\zeta(A,\mathscr{P}) = \zeta(B,\mathscr{Q})$ *then* $(A,\mathscr{P}) \sim (B,\mathscr{Q})$.

Proof. It is well know that since μ is non-atomic, there exists $\{W_1,\dots W_n\} \in \texttt{part}(\mathcal{U})$, such that $\mu(W_i) = p_i$ for $i \leq n$ (for example, see Billingsley (1995) Problem 2.19(d)). By richness, there exists some $(A,\{P_1,\dots,P_m\})$ such that $\{W_{A,P_i}\}_{i\in m}$ is μ-equivalent to $\{W_1,\dots W_n\}$. By Lemma 2, it is without loss of generality to assume $\mu(W_{A,P_i}) = \mu(W_i)$ for $i \leq n$; it follows that $\mu(W_{A,P_j}) = 0$ for $j > n$. Then $(A,\{P_1 \cup \bigcup_{j>n} P_j, P_2,\dots P_n\}$ is the desired experiment. The later claim follows directly from **A7**. ★

In light of Lemma 9, we can define the functional $\eta : \texttt{prob}^* \to \mathbb{R}$ via

$$\eta(p_1,\dots,p_n) = \texttt{vnm}(A,\mathscr{P}),$$

where $(A,\mathscr{P}) \in \zeta^{-1}(p_1,\dots,p_n)$. Extend η to all of $\texttt{prob}$ by simply ignoring 0s. That is, for each $(p_1,\dots,p_n) \in \texttt{prob}$, set $\eta(p_1,\dots,p_n) = \eta(p_{k_1},\dots,p_{k_m})$, where $k_1,\dots,k_m \subseteq 1,\dots,n$ is the subsequence that selects strictly positive entries. Our normalization $\texttt{vnm}(\{x\},\{\{x\}\}) = 0$ implies $\eta(1) = 0$.

For $p = (p_1,\dots p_n) \in \texttt{prob}^*$ and $\{q^i\}_{i\leq n}$, where each $q^i = (q^i_1,\dots,q^i_{m^i}) \in \texttt{prob}^*$ is of (possibly distinct) length m^i, let

$$p \otimes \{q^i\}_{i\leq n} = (p_1q^1_1,\dots,p_1q^1_{m^1},\dots,p_nq^n_1,\dots,p_nq^n_{m^n}).$$

We can intuitively view $p \otimes \{q^i\}_{i \leq n}$ as the reduction of a compound lottery over over $\sum_{i \leq n} m^i$ outcomes, thinking of p as the marginal on n first stage lotteries and q^i the conditional lottery on m^i outcomes given the realization of p.

We will now show that η satisfies the following three properties:

(K1) $\eta(p_1, \ldots, p_n) = \eta(p_1, \ldots, p_n, 0)$

(K2) $\eta(p_1, \ldots, p_n) \leq \eta(\frac{1}{n}, \ldots, \frac{1}{n})$

(K3) $\eta(p \otimes \{q^i\}_{i \leq n}) = \eta(p) + \sum_{i \leq n} p_i \eta(q^i)$

Property (K1) follows immediately from the construction of η, in particular how it is extended from $\texttt{prob}^*$ to $\texttt{prob}$. (K2) follows immediately from **A7**. We will show (K3).

Fix some $p \in \texttt{prob}^*$ and $\{q^i\}_{i \leq n}$, with each $q^i \in \texttt{prob}^*$. For each $0 \leq k \leq n$, let

$$q^{i,k} = \begin{cases} q^i & \text{if } i \leq k \\ (1) & \text{otherwise} \end{cases}$$

Notice that $q^{i,n} = q^i$ and $p \otimes \{q^{i,0}\} = p$. Thus, the result follows by showing that

$$\eta(p \otimes \{q^{i,k}\}) = \eta(p \otimes \{q^{i,k-1}\}) + p_k \eta(q^k), \tag{11}$$

for $0 < k \leq n$.

For for $1 \leq i \leq n$ and $1 \leq j \leq m^i$, set $r^i_j = p_i q^i_j$. With this notation we can write the relevant distributions as

$$p \otimes \{q^{i,k-1}\} = (r^1_1, \ldots, r^1_{m^1}, \ldots, r^{k-1}_1, \ldots, r^{k-1}_{m^{k-1}}, p_k, p_{k+1}, \ldots p_n)$$
$$p \otimes \{q^{i,k}\} = (r^1_1, \ldots, r^1_{m^1}, \ldots, r^k_1, \ldots, r^k_{m^k}, p_{k+1}, \ldots p_n)$$

From Lemma 9, we obtain some $e^k = (A, \{R^1_1, \ldots, R^k_{m^k}, P_{k+1}, \ldots, P_n\})$ in $\zeta^{-1}(p \otimes \{q^{i,k}\})$ and also some $e' = (B, \{Q_1, \ldots, Q_{m^k}\})$ in $\zeta^{-1}(q^k)$. Define $P_k = \bigcup_{j=1}^{m^k} R^k_j$. By construction, $\mu(W_{A,P_k}) = \sum_{j=1}^{m^k} \mu(W_{A,R^k_j}) = \sum_{j=1}^{m^k} p_k q^k_j = p_k$. Thus, we have $e^{k-1} = (A, \{R^1_1, \ldots R^{k-1}_{m^{k-1}}, P_k, \ldots, P_n\})$ is in $\zeta^{-1}(p \otimes \{q^{i,k-1}\})$. Using the

definition of η, we have

$$\begin{aligned}\eta(p \otimes \{q^{i,k-1}\}) &= \texttt{vnm}(e^{k-1}) \\ \eta(p \otimes \{q^{i,k}\}) &= \texttt{vnm}(e^{k}) \\ \eta(q^j) &= \texttt{vnm}(e')\end{aligned} \tag{12}$$

At last, we have the requisite ingredients to appeal to **A8**, and the representation via vnm, obtaining

$$\tfrac{1}{1+\mu(W_{A,P_k})}\texttt{vnm}(e^k) + \tfrac{\mu(W_{A,P_k})}{1+\mu(W_{A,P_k})}0 = \tfrac{1}{1+\mu(W_{A,P_k})}\texttt{vnm}(e^{k-1}) + \tfrac{mu(W_{A,P_k})}{1+\mu(W_{A,P_k}))}\texttt{vnm}(e')$$

Simplifying and plugging in the suitable replacements via (12) yields the desired relation.

Theorem 1 of Khinchin (1957) shows that if η satisfies the properties (K1)–(K3), it take the form

$$\eta(p_1, \ldots, p_n) = -\lambda \sum_{i=1}^{n} p_i \log(p_i), \tag{13}$$

where $\lambda > 0$.[10] Since we are free to rescale an expected utility representation by a positive constant (we only used a single degree of freedom in choosing the intercept $\texttt{vnm}(\{x\}, \{\{x\}\}) = 0$) we can set $\lambda = 1$.

Finally, let $(A, \{P_1, \ldots, P_n\}) \in \mathbb{E}$. Without loss of generality, assume the first $k \leq n$ observations have positive probability (i.e., $\mu(W_{A,P_i}) > 0$ if and only if $i \leq k$). Set $P^\dagger = P_1 \cup \bigcup_{i=k+1}^{n} P_i$ We have

$$\begin{aligned}\texttt{vnm}(A, \{P_1, \ldots, P_n\}) &= \texttt{vnm}(A, \{P^\dagger, P_2, \ldots, P_k\}) && \text{(from \textbf{A6})} \\ &= \eta(\mu(W_{A,P_1}), \ldots, \mu(W_{A,P_k})) && \text{(definition of } \eta) \\ &= \eta(\mu(W_{A,P_1}), \ldots, \mu(W_{A,P_n})) && \text{(from (K1))} \\ &= -\sum_{i \leq n} \log(\mu(W_{A,P_i}))\mu(W_{A,P_i}), && \text{(from (13))}\end{aligned}$$

[10] The property (K3) in Khinchin (1957) is stated slightly differently: it requires all q^i to be the same length, but allows for zero-probability entries. These formulations are clearly equivalent under (K1), where 0s can be added to make each q^i the same length.

as needed to complete the proof. ■

References

Billingsley, Patrick, "Probability and measure. 1995," *John Wiley&Sons, New York*, 1995.

Chaloner, Kathryn and Isabella Verdinelli, "Bayesian experimental design: A review," *Statistical science*, 1995, pp. 273–304.

Cover, Thomas M, Joy A Thomas et al., "Entropy, relative entropy and mutual information," *Elements of information theory*, 1991, *2* (1), 12–13.

Dekel, Eddie, Barton L Lipman, and Aldo Rustichini, "Representing preferences with a unique subjective state space," *Econometrica*, 2001, *69* (4), 891–934.

Drake, Marshall, Fernando Payró, Neil Thakral, and Linh T Tô, "Bayesian adaptive choice experiments," Technical Report, Mimeo 2022.

Ergin, Haluk and Todd Sarver, "Hidden actions and preferences for timing of resolution of uncertainty," *Theoretical Economics*, 2015, *10* (2), 489–541.

Gilboa, Itzhak and Ehud Lehrer, "The value of information-An axiomatic approach," *Journal of Mathematical Economics*, 1991, *20* (5), 443–459.

Grnbaum, Branko, *Convex Polytopes, volume 221 of Graduate Texts in Mathematics* 2003.

Gul, Faruk and Wolfgang Pesendorfer, "Temptation and self-control," *Econometrica*, 2001, *69* (6), 1403–1435.

_ and _ , "Random expected utility," *Econometrica*, 2006, *74* (1), 121–146.

Khinchin, A. I., *Mathematical foundations of information theory*, Vol. 434, Courier Corporation, 1957.

Lindley, Dennis Victor, *Bayesian statistics: A review*, SIAM, 1972.

Louviere, Jordan J and David A Hensher, "On the design and analysis of simulated choice or allocation experiments in travel choice modelling," *Transportation research record*, 1982, *890* (1), 11–17.

Luce, R Duncan, "On the possible psychophysical laws.," *Psychological review*, 1959, *66* (2), 81.

McFadden, Dennis, "Precedence effects and auditory cells with long characteristic delays," *The Journal of the Acoustical Society of America*, 1973, *54* (2), 528–530.

Raiffa, Howard and Robert Schlaifer, *Applied statistical decision theory*, Vol. 78, John Wiley & Sons, 2000.

Rao, KPS Bhaskara and M Bhaskara Rao, *Theory of charges: a study of finitely additive measures*, Academic Press, 1983.

Rockafellar, R Tyrrell, *Convex Analysis*, Vol. 28, Princeton University Press, 1970.

Stovall, John E, "Temptation with uncertain normative preference," *Theoretical Economics*, 2018, *13* (1), 145–174.