Representation Theorems

Mark Dean

Lecture Notes for Fall 2018 PhD Class in Behavioral Economics - Columbia University

1 Introduction

A key set of tools that we are going to make use of in this course are 'representation theorems'. These are theorems of the following type:

> <**Some Data**> is consistent with <**some model**> consisting of <**some model elements**>if and only if it satisfies <**some axioms**>

Representation theorems are going to be useful for us because many of the models we want to work with are couched in terms of 'latent variables' - i.e. variables that are not directly observable (utility, beliefs, attention costs etc). Representation theorems provide us with a way to test such models/

2 Two Familiar Representation Theorems

To discuss what representation theorems are, and why they are so powerful, it is going to be useful to have some concrete examples to work with.

2.1 Example 1: Choice and Preference Maximization

Our first example is going to ask the following question: under what circumstances can we think of a decision maker (DM) as a preference maximizer? In other words, when can the choices of a

DM be represented as resulting from the maximization of a complete, transitive, reflexive binary relation on X? This is an example that you should have come across before, and is an important one, as almost all economics builds on the assumption that people maximize stable, well behaved preferences.

In order to answer this question, we are going to prove a representation theorem: we are going to write down a set of conditions, or axioms on the choices people make, and show that these axioms are equivalent to the statement that people are preference maximizers: if a DM satisfies these axioms then we can think of them as a preference maximizer. If they don't then we cannot. In order to do this, we are going to have to be more formal about what we mean by the various parts of the above sentence.

- First of all, choices over what? We are going to think of choices from subsets of a grand set X. Initially, we will assume that X is finite, and that an object in the set is just an object - it doesn't have any other characteristics. They could be fruit, musical instruments, stocks, infinite consumption streams, lotteries or anything else. All we are going to know about each object is the label that identifies it in the set. We will relax both these assumptions later in the course.

- What do we mean by choices? For now, let us imagine that we observe the choices that our DM makes from every subset of X. We will represent these choices by a *complete choice correspondence* $C : 2^X/\varnothing \to 2^X/\varnothing$ such that $C(A) \subset A$ for all $A \in 2^X/\varnothing$. Note that here we are allowing the decision maker to choose more than one option from any given choice set. This is a technically useful assumption (because it allows us to deal with indifference), but an observationally very dubious one - a point that we will come back to later. Notice also that we are assuming that we observe our DM choose once (and only once) from each subset of X. This is also a strong assumption, and one that we will again relax later.

- What do we mean by a preference maximizer? Formally, a binary relation B is a subset of $X \times X$. We will use the symbol $\succeq$ to represent our binary relation, so $x \succeq y$ means that $\{x, y\} \in B$. We demand that our decision maker behaves according to a binary relation that has certain properties:

  Definition 1 *A transitive binary relation is one in which $x \succeq y \succeq z$ implies $x \succeq z$. A reflexive binary relation is one such that $x \succeq x$ for all $x \in X$. A binary relation that is*

transitive and reflexive is called a preference relation (or a preorder). A complete binary relation is one for which, for any $x, y \in X$, either $x \succeq y$ or $y \succeq x$ or both.

We will think of the complete preference relation as representing 'weak preferences' - i.e. $x \succeq y$ means that 'x is at least as good as y'. Our behavioral model is that people have preferences that are well behaved (in the sense of being complete, transitive and reflexive), and these preferences govern their choices. (It will become obvious why we think of such preferences as well behaved).

Our question is, under what circumstances can we find **some** complete preference relation such that choice is equal to the set of maximum objects in each set according to that preference ordering: the DM chooses the best objects according to that preference ordering. In other words, we want to find some $\succeq$ such that, for all $A \in 2^X/\varnothing$

$$C(A) = \{x \in A | x \succeq y \ \forall \ y \in A\}$$

Note that we are allowing for the possibility that two objects are indifferent - that $x \succeq y$ and $y \succeq x$. In this case, and if both objects are preferred to all other objects in some choice set, then we want both objects to be 'chosen' (in some not very well defined sense).

We can now define our problem more formally. The aim of our representation theorem is to find some conditions on the choice function C such that we can find some preference relation that *rationalizes* the DM's choices (i.e. the DM chooses the best objects according to those preferences). Note here that the concept of observability is crucially important. We assume that we **can** observe choices but **cannot** observe preferences. If we could observe preferences then it would be easy to test whether people are maximizing preferences - all we would have to do is look and see whether the item they chose in each set was the most preferred item. Instead, we are completely agnostic about what this preference relation is, we just want the DM to be behaving in a manner consistent with *some* preference relation.

So what are the relevant conditions? At this point, assuming you have taken the graduate micro class, you should be screaming 'WARP'![1]. And you would be right. However, it is going to

[1] The Weak Axiom of Revealed Preference (WARP) states:

Axiom 1 (WARP) *If* $\{x, y\} \in A \cap B$, $x \in C(A)$, *and* $y \in C(B)$ *then* $x \in C(B)$

be convenient for us to break WARP down into two pieces (as originally done by Armartya Sen):

Axiom 2 (Property α) *If $x \in B \subseteq A$ and $x \in C(A)$, then $x \in C(B)$*

Axiom 3 (Property β) *If $x, y \in C(A)$, $A \subseteq B$ and $y \in C(B)$ then $x \in C(B)$*

It is worth stopping and thinking for a minute about these two axioms. The first is equivalent to the independence of irrelevant alternatives, and is very intuitively appealing. It says that, if you choose an alternative x from a larger set, then take some objects out of that set that are not x, then you should still choose x from the smaller set. This is clearly a property you would expect a 'rational' decision maker to obey - if they choose x from the larger set, they are telling you that they prefer x to all of the other objects in that set. They should therefore prefer x to any objects in a *subset* of that larger set.

What about the second property? The first thing to note is that this property only has bite in the case of choice *correspondences*. If C is single valued, then the condition $x, y \in C(A)$ can never hold, so this axiom will be satisfied trivially. In the case of a choice correspondence, this condition says that, if x and y are chosen from a set, and y is chosen from a superset of that set, then x must also be chosen from that superset. Again, this makes sense if we think of a rational decision maker. If x and y are chosen together, then the DM must be indifferent between them. If y is chosen from some other set, then it must be at least as good as anything else in that set, and therefore so must x. (Note, why does property β restrict itself to B's which are supersets of A? Surely this property should hold for any B? What is going on here?)

You should convince yourself that, properties α and β between then are equivalent to WARP.

We are now in a position to state and prove our representation theorem. You are probably familiar with the proof, but we will go through it again in order to highlight some points

Definition 2 *A choice correspondence $C : S \to 2^X/\varnothing$ for some $S \subset 2^X/\varnothing$ is **rationalized** by a preference relation $\succeq$ if, for every $A \in 2^X/\varnothing$, it is the case that*

$$C(A) = \{x \in A | x \succeq y \ \forall \ y \in A\}$$

Theorem 1 *For any finite set X and complete choice correspondence $C : 2^X/\varnothing \to 2^X/\varnothing$, there*

exists a complete preference relation $\succeq$ that rationalizes that choice correspondence if and only if C satisfies property α and β.

Proof. *The first thing to do is note that this proof must come in two parts, as we are making two claims: this comes from the fact that the statement is "if and only if', so we have to show (i) that α and β imply that we can find a rationalizing preference relation and (ii) any rationalizable choice function satisfies α and β. We will start with the former, as this is the more tricky bit (in fact, we have already argued informally for the latter.)* ■

Proof (axioms imply representation). *We will break the proof down into the following steps*

1. ***Generate a candidate binary relation.*** *Our claim is that, if the choice correspondence satisfies α and β, then it is rationalizable by some complete preference relation. The first stage of the proof is to describe such a relation, which we will then shows does the necessary job. We will define the relationship using choices from two objects we will say that $x \trianglerighteq y$ if and only if $x \in C(\{x, y\})$, so x is 'weakly preferred' to y (according to our candidate preference relation) if it is chosen from the set containing x and y only. We will stretch this definition somewhat by saying that $x \trianglerighteq x$, as x is definitionally chosen from the set $\{x\}$.*

2. ***Show that $\trianglerighteq$ is a complete preference relation.*** *So we have defined a binary relation. Great. However, our theorem demands that choices be rationalized by a complete preference relation - i.e. a complete, transitive, reflexive binary relation. We next need to show that $\trianglerighteq$ has these properties. Reflexivity is easy - in fact we defined $\trianglerighteq$ explicitly so that it is reflexive. Completeness is also relatively straightforward. By definition, $C(\{x, y\})$ is either $\{x\}$, $\{y\}$ or $\{x, y\}$. Thus, by the construction of $\trianglerighteq$ either $x \trianglerighteq y$, $y \trianglerighteq x$ or both. Finally, we need to show transitivity, which we will do by contradiction. Imagine there exists $x, y, z \in X$ such that $x \trianglerighteq y \trianglerighteq z$ but not $x \trianglerighteq z$. This implies*

$$\begin{aligned} x &\in C(\{x, y\}) \\ y &\in C(\{y, z\}) \\ x &\notin C(\{x, z\}) \end{aligned}$$

This in turn implies that $z \in C(\{x, z\})$. We can now show that we must have a violation of either property α or property β. Consider the set $\{x, y, z\}$. If $x \in C(\{x, y, z\})$, then the fact that $x \notin C(\{x, z\})$ is a direct violation of property α. If $y \in C(\{x, y, z\})$, then by property α,

$y \in C(\{x,y\}) = \{x,y\}$. *Property* β *then implies that* $x \in C(\{x,y,z\})$, *which we have already shown leads to a violation of* α. *If* $z \in C(\{x,y,z\})$, *then by* α $z \in C(\{y,z\}) = \{x,z\}$, *and so by* β $y \in C(\{x,y,z\})$. *Again, we have already shown that this leads to a violation. However, as* $C(\{x,y,z\})$ *is nonempty, one of these cases must occur, and so a failure of transitivity implies a failure of either* α *or* β.

3. ***Show that*** $\trianglerighteq$ ***rationalizes*** C. *We now need to show that, for all sets, our DM chooses as if they are maximizing* $\trianglerighteq$. *In other words, for some arbitrary* $A \in 2^X/\varnothing$ *we need to show that* $C(A) = \{x \in A | x \trianglerighteq y \ \forall \ y \in A\}$. *As we are proving the equality of two sets, this in itself takes two stages*

   (a) $C(A) \subseteq \{x \in A | x \trianglerighteq y \ \forall \ y \in A\}$. *Say* $x \in C(A)$. *Take any* $y \in A$. *We need to show that* $x \trianglerighteq y$ *- in other words that* $x \in C(\{x,y\})$. *However, this follows directly from property* α. *Thus, anything that is chosen from* A *must be 'preferred' to everything else in* A

   (b) $C(A) \supseteq \{x \in A | x \trianglerighteq y \ \forall \ y \in A\}$. *Say* $x \trianglerighteq y \ \forall \ y \in A$. *Then,* $x \in C(\{x,y\})$ *for all* $y \in A$. *Now* $C(A)$ *must be non-empty, so either* $x \in C(A)$ *(in which case we are done), or* $y \in C(A)$ *for* $y \neq x$. *By property* α, *this implies that* $\{x,y\} = C(\{x,y\})$, *and so by property* β, $x \in C(\{x,y\})$

This shows that properties α *and* β *are sufficient for rationalizability* ■

Proof (representation implies axioms). *The next thing that we have to do is show the 'only if' part of the statement - that rationalizability implies properties* α *and* β. *In other words, we have to show that if there is a complete preference relation* $\succeq$ *such that* $C(A) = \{x \in A | x \succeq y \ \forall \ y \in A\}$, *then* C *must obey properties* α *and* β

- *Property* α*:* $x \in C(A)$ *implies that* $x \succeq y \ \forall \ y \in A$. *Thus* $x \succeq y \ \forall \ y \in B$ *for any* $B \subseteq A$. *Thus, as,* $\succeq$ *rationalizes choice from* B, *then it must be the case that* $x \in C(B)$
- *Property* β*: If* $x, y \in C(A)$, *then* $x \succeq y$. *If* $y \in C(B)$, *then* $y \succeq z \ \forall \ z \in B$. *Thus, by transitivity,* $x \succeq z \ \forall \ z \in B$, *and, as* $\succeq$ *rationalizes choice in* B, $x \in C(B)$

■

This completes the proof of our first representation theorem.

One final thing to note. The preference relation that rationalizes a complete set of choice data is unique. For completeness we will prove this claim as well:

Theorem 2 *Let C be a choice correspondence that satisfies properties α and β. There is one and only one preference relation that rationalizes C*

Proof. *The fact that there is such a preference relation we have already proved. We will prove uniqueness by contradiction. Say $\succeq_1$and $\succeq_2$ both rationalized C ,and $\succeq_1 \neq \succeq_2$. Without loss of generality, this implies that there exists an x and y such that $x \succeq_1 y$ but not $x \succeq_2 y$. But the former statement implies that $x \in C(\{x, y\})$ while the latter implies $x \notin C(\{x, y\})$, a contradiction.* ■

2.2 Example 2: Preferences and Utility Maximization

The second example we are going to work through is another that you should be thoroughly familiar with: Under what circumstances is it possible to represent 'preferences' with a numerical utility function (note that the language here has become a bit difficult as we have defined a 'preference relation' already. What we really mean is: under what circumstances can we represent a binary relation numerically). To put matters more formally, we want to find a *utility representation* for a binary relation

Definition 3 *A binary relation $\succeq$ on a set X has a utility representation if there exists a utility function $u{:}X \to \mathbb{R}$ such that*

$$\begin{aligned} u(x) &\geq u(y) \\ \text{if and only if } x &\succeq y \end{aligned}$$

for all x, $y \in X$

Again, this is a pretty fundamental question, as almost all of economics uses utility functions, rather than preference relations as their basis. This is because we have a load of cool tools to work with utility functions and not very many cool tools to work with binary relations (though note that most of these cool tools require the utility function to be differentiable - something that we will not say anything about at the moment). It is also one that you almost certainly already

know the answer to - the properties that we require (if X is finite) are completeness, reflexivity and transitivity. Note that it is no coincidence that these are the properties that we used to define 'well behaved' preferences.

Before we proceed - note that we have changed our assumptions about observability. Here, we are assuming that preferences *are* observable, but utility is not. We shall come back to this point later.

Theorem 3 *Let X be a finite set. A binary relation $\succeq$ on X has a utility representation if and only if is a complete preference relation.*

Proof. *Again, we have two things to prove here, as this is an if and only if statement. Again, we will begin by showing that the axioms imply the representation, which is the more difficult direction.* ■

Proof (axioms imply representation). *We will proceed using induction on the size of the set X. That is, we will show that (i) it is true for $|X| = 1$ and (ii) if it is true for $|X| = n - 1$ then it is true for $|X| = n$. The case of $|X| = 1$ is trivial (though note that it uses reflexivity), so we will move onto the second part of the proof. Let X be a set of size n, and let $\succeq$ be a complete preference relation on x. Remove object from the set X, which we will denote x^*. Now note that X/x^* is a set of size $n - 1$ and $\succeq$ induces a complete preference relation on X/x^* (yes?). Thus, there is a function $v : X/x^* \to \mathbb{R}$ such that $v(x) \geq v(y)$ if and only if $x \succeq y$. We will use this to construct a utility function u on X. We will set $u(x) = v(x)$ for all $x \in X/x^*$. This utility function will clearly represent $\succeq$ on X/x^* in the sense that $u(x) \geq u(y)$ if and only if $x \succeq y$ for all $x, y \in X/x^*$. Thus, all that remains to do is to is to set $u(x^*)$ and show that the utility function works here to. There are 4 cases.*

1. *$x^* \succeq \bar{x}$ and $\bar{x} \succeq x^*$ for some $\bar{x} \in X/x^*$. In this case, we set $u(x^*) = u(\bar{x})$. Now, note that, for any $y \neq x^*$*

$$\begin{aligned} u(x^*) &\geq u(y) \\ \text{if and only if } u(\bar{x}) &\geq u(y) \\ \text{if and only if } \bar{x} &\succeq y \\ \text{if and only if } x^* &\succeq y \end{aligned}$$

The third line follows from the fact that $\bar{x}$ and $y \in \bar{x} \in X/x^*$, and so by the inductive hypothesis u represents the relationship between these two. The last line follows from transitivity. Using the same technique it is possible to show that $u(y) \geq u(x^*)$ if and only if $y \succeq x^*$

2. *$x^* \succeq y$ for all $y \in X$. (for the next three cases we will assume that there is no $\bar{x} \in X/x^*$ such that $x^* \succeq \bar{x}$ and $\bar{x} \succeq x^*$). In this case we set*

$$u(x^*) = \max_{y \in X/x^*} v(y) + 1$$

Now, for any $y \neq x$, $u(x^*) > u(y)$ and $x^* \succeq y$. By assumption $y \succeq x^*$ for no $y \neq x^*$

3. *$y \succeq x^*$ for all $y \in X$. In this case, we set*

$$u(x^*) = \min_{y \in X/x^*} v(y) - 1$$

Now, for any $y \neq x$, $u(y) > u(x^*)$ and $y \succeq x$. By assumption $x^* \succeq y$ for no $y \neq x^*$

4. *There exists at least one $y \in X/x^*$ such that $y \succeq x$ and $z \in X/x^*$ such that $x \succeq z$. In this case, define two sets: $X^* = \{x \in X/x^* | x \succeq x^*\}$ and $X_* = \{x \in X/x^* | x^* \succeq x\}$. Note that these two sets are disjoint (as we have ruled out the possibility that $x \succeq x^*$ and $x^* \succeq x$ for any $x \neq x^*$), and that, for any $x \in X^*$ and $y \in X_*$, $x \succeq y$ but not $y \succeq x$ ($x \succeq y$ follows directly from transitivity. If $y \succeq x$, then $x^* \succeq y \succeq x^*$, which we have ruled out by assumption). This in turn implies that*

$$\min_{x \in x^*} v(x) > \max_{y \in X_*} v(y)$$

We will therefore set

$$u(x^*) = \frac{1}{2} \min_{x \in x^*} v(x) + \frac{1}{2} \max_{y \in X_*} v(y)$$

Thus, for any $x \neq x^*$

$$\begin{aligned} u(x^*) &\geq u(x) \\ \text{if and only if } x &\in X_* \\ \text{if and only if } x^* &\succeq x \end{aligned}$$

Similarly

$$\begin{aligned} u(x) &\geq u(x^*) \\ \textit{if and only if } x &\in X^* \\ \textit{if and only if } x &\succeq x^* \end{aligned}$$

This completes the first part of the proof. ■

Proof (Representation Implies Axioms). *This direction is relatively simple. Say that* $\succeq$ *is a binary relation on* X *and that* $u : X \to \mathbb{R}$ *is a utility representation of that function. Then* $u(x) \geq u(x)$ *implies* $x \succeq x$ *(reflexivity), for any* x, y *either* $u(x) \geq u(y)$ *or* $u(y) \geq u(x)$ *implying either* $x \succeq y$ *or* $y \succeq x$ *(completeness), and that* $x \succeq y \succeq z$ *implies* $u(x) \geq u(y) \geq u(z)$*, and so* $x \succeq z$ *(transitivity).* ■

Finally, note that the utility function that can represent a complete preference relation is *not* unique. It is unique only up to strictly increasing transformation. This means that, if the function u represents a set of preferences, then the function v will represent the same preferences if and only if v is a strictly increasing transform of u.

Theorem 4 *Let* $u : X \to \mathbb{R}$ *be a utility representation for a complete preference relation* $\succeq$*. Then* $v : X \to \mathbb{R}$ *will also represent* $\succeq$ *if and only if there is a strictly increasing function* T *such that*

$$v(x) = T(u(x)) \ \forall \ x \in X$$

Proof. *To show the if part, note that, if* v *is a strictly increasing transform of* u *then*

$$\begin{aligned} v(x) &\geq v(y) \\ \textit{if and only if } u(x) &\geq u(y) \\ \textit{if and only if } x &\succeq y \end{aligned}$$

To show the only if part, note that if v *is not a strictly increasing transform of* u*, then there exists an* x *and* y *such that* $u(x) > u(y)$ *but* $v(x) \leq v(y)$*.* $u(x) > u(y)$ *implies that it is not the case that* $y \succeq x$*. Thus,* v *does represent* $\succeq$*.* ■

This uniqueness result is important, as it tells us how much information is in the utility function. In this case, it is telling us that it is only the ordinal (ordering) information that is important - that the utility number is bigger than another. The magnitude of those differences are meaningless. It is therefore meaningless to say things like 'the utility of x is twice that of y', because we could just as well use another utility function in which the utility of x is a million times that of y, or one where the utility of x is 1% higher than that of y. Any utility function that preserves the same ordering properties will do the job.

3 What Have We Just Done?

These are two important representation theorems, and the proofs contain some of the tricks that you will see again in more complicated settings. However, they are also theorems that you have probably come across before. One of the reasons that I wanted to put them on the table is so that we can think about the structure that these theorems have in common. This is also going to allow us to discuss two different philosophical approaches to decision theory. Both of these approaches have the same first stage:

Stage 1: Define the primitive 'data set'. The first job in constructing a representation theorem is to think about the properties of the observations to which you are going to apply your axioms.

In the first case above, we took as our observations the choices made by the DM from different sets of objects. We assumed that the objects were just objects - they didn't have any other characteristics. This is an assumption we will change later on. Instead, we may assume that the objects of choice are probability distributions (lotteries), or consumption streams or bundles of goods. We assumed that there was only a finite number of them. We also assumed that we observed DMs choose once from every subset of a grand choice set, but only once from each choice set. Finally, we assumed that we could observe choice correspondences, rather than just choice functions. All these assumptions played an important role in the nature of the representation theorem that we eventually ended up with.

In the second example, the primitive of the representation function was the preferences of the DM, rather than choices. This might seem a little puzzling - in what way can we think of

preferences as data? This is an issue that we will come back to. However, note again that we have choices to make - for example again about the properties of the objects over which the preferences were defined.

In most cases, the primitive data sets that decision theorists deal with either come in the form of choices or preferences. However, there are many different variations. For example, one might take as one's primitive observation of choices from a choice set and a reference point (i.e. rather than $C(A)$, we consider $C(A, z)$, the choice from A when the reference point is z.) Or we might consider choices from a choice set after the decision maker has thought about the problem for a certain length of time (i.e $C(A, t)$, the choice from A after the DM has thought about the problem for length of time t). We will consider both of these examples during the course. One could also consider data sets of a completely different nature - for example one model we will consider is on in which the data set is a function $\delta(z, p)$, which we interpret as the amount of the neurotransmitter dopamine released when a prize z is obtained from a lottery p. Any and all of these data sets are amenable to 'decision theoretic' analysis.

What is the next stage of the theory? Well, at the end of the day, we are going to end up with a representation theorem linking a set of axioms concerning the data set to a model of that data set. The question is, which of these comes first? Do you start off by thinking of a set of intuitively plausible axioms, and then show that these axioms imply some model of behavior in that data set? Or do you start off with a model of what is going on in the data set, then find a set of axioms that capture the behavioral implications of that model? I think that the traditional approach has been the former: axioms come first, with the model being derived from those axioms. However, in my view, the most useful way to use decision theory is the latter - it allows you to say something concrete about the observable implications of your model. We will come back to this point below. For now we shall simply note the two different approaches.

Stage 2 Version 1 - The Traditional Approach: Define a set of axioms These are a set of statements concerning the data set. In the first case above our axioms were properties α and β - simple, testable and intuitively plausible statements about how people make choices. In the second, our axioms were completeness, transitivity and reflexivity. Again, easily testable and intuitively plausible statements about the primitive data set (in this case preferences).

Stage 2 Version 2 - The Alternative Approach: Define a behavioral model. The alter-

native approach is to next think about a plausible model to explain what is going on in our data set. In our first example, the model is that people are making choices in order to maximize some well behaved preference relation. Notice that here we assume that preferences are *unobservable,* so it is not immediately obvious how to test this model using our data set. In the second case, our model is that people have preferences that are derived from the process of utility maximization. Again, in this case we are assuming that utility is not directly observable (though preferences are), meaning that the observable implication of this model are once again unclear.

Stage 3: The representation theorem. The next stage, whether one is coming from the traditional or the alternative approach is to prove a representation theorem. This is a theorem that links together a set of axioms to a behavioral model. In our first example, we showed that preference maximization is equivalent to properties α and β. In the second, we showed that completeness, transitivity and reflexivity were the same as the existence of a utility representation. Note that, in both cases, these theorems are 'if and only if' - the axioms are necessary and sufficient for the representation. This is the gold standard of these theories - it means that the axioms are *exactly equivalent* to the behavioral model. In our first example, if properties α and β hold then there is some set of preferences that rationalize the choice data. If the do not, there is no such preference relation. Thus, α and β are the exact observable implications of the model of preference maximization.

Stage 4: Uniqueness result. In both our examples, we finished by proving a uniqueness result. In the first case we showed that there was only one preference relation that could rationalize a set of choices. In the second case we showed that there were many utility functions that could represent a set of preferences, but they would all be linked by strictly increasing transformations. This is an important step of the process, as it tells us how 'seriously' to take our representation. In our second example, we should take the ordinal information in the utility function very seriously, but not the cardinal information.

4 Why Was This A Good Idea?

So now we have established what a representation theorem is: an equivalence result between a set of observable axioms and a representation - a behavioral model that may rely on unobservable

elements. We also know what the 4 steps that are involved in developing a representation theorem. What we have yet to cover is *why* this is an interesting thing to do. Again, we are going to discuss two approaches The first, which I described as the 'traditional approach' above, sees the axioms themselves as interest. This could be for a number of reasons. - they could be seen as 'self evidently true' (i.e. axiomatic in the standard meaning of the word). They could be seen as justifiable as the definition of rational behavior (e.g. someone who has intransitive preferences is *by definition* being irrational). They could be justified as capturing an essential element of a certain type of behavior (Gul and Pessendorfer took this approach when they thought about temptation and self control. They posited that the essential behavioral characteristic that defined temptation and self control is that a person with self control issues may sometimes choose to restrict their own choice sets - i.e. choose to have a smaller, rather than a larger choice set. This property they formalized as the *set betweenness* axiom - allowing for the preference of smaller over larger choice sets in a structured way.)

For all these reasons, people may be interested in the axioms that govern behavior. But this does not explain why they might be interested in representation theorems. Why do they care what sort of models are equivalent to these axioms? I think that there are a few reasons. One is that, if (for example) you do believe that these axioms are capturing rational choice, then you might be interested in what behavioral models are 'rational'. We have shown that utility maximization is 'rational', but it turns out that other models will also lead to 'rational' outcomes. A second reason is more practical - choice functions and binary relations are not very easy to work with, while utility functions are! We have all sorts of mathematical tricks that we can use to find the maximum of a utility function (i.e. the toolkit of static optimization) that just don't work on binary relations. Thus, if we believe that people are preference maximizers, it is very useful to know that we can treat them as utility maximizers (though to use the power of our static optimization toolkit we also need the utility function to be differentiable, which is not guaranteed by anything we have done so far, not least because we have only covered the case of a finite choice set.) Similarly, it is useful to know that my behavioral assumptions allow me to model people as expected utility maximizers, exponential discounters etc. (though, again, to use these tricks, we need to know something about the differentiability of the resulting utility functions, something we have yet to say anything about).

While I can see some of the power of these arguments, this does not, in general, represent the way that I use decision theory. Instead, I tend to go in the other direction. I *start* with

some behavioral model of how people make decisions, and use decision theory to understand the observable implications of this model. In other words, I have some intuition about what sensible decision making procedures, but I want some way to test whether this intuition is right. For example, I may think that preference maximization sounds like a reasonable model of behavior, but as a (social) scientist I would like to be able to test this. Due to the above result, I know that means testing properties α and β.

Why do I need to go through the rigmarole of a representation in order to find the testable implications? Because the models that we use to describe decision making tend to have unobservable elements. Take for example the model of utility maximization. If I looked at objects and saw their utility, then I wouldn't *need* a representation theorem in order to derive testable implications - I would just look and see whether people did in fact choose the highest utility object in each case (consider a model of choice over amounts of money, where we assume that people choose more money to less - we would not need a representation theorem to test this model! Or rather, any such theorem would be trivial). However, because utility is not observable, it is not immediately obvious how to test whether there is some utility function that rationalizes the data - i.e. that people are acting as if they are utility maximizers.

Is there an alternative? Yes, in fact there is. Take the model of utility maximization. Rather than ask the question of whether there is *some* utility function that can rationalize our DMs choices, we could make some assumptions about what that utility function should look like. For example, if the objects of choice were lotteries over money, we could assume that people had a constant relative risk aversion utility function, estimate the parameters of this utility function, and then see how well the resulting estimated utility function explains the DM's choices. In fact, this approach is taken a lot by economists. What are the disadvantages of this approach? In my opinion there are a few[2] but here I would like to highlight two.

1. The resulting test is now a joint test of two hypotheses: that people maximize utility and that utility is of the functional form that you have assumed. Moreover, if you think of more general objects (such as teapots), it becomes very difficult to see how one could sensibly come up with a model that assigned utility based on the properties of that object (length of spout?). In fact, there were lots of articles in early economics with titles such as 'The Seven

[2]Which I highlight in "Axiomatic Methods, Dopamine and Reward Prediction Error" (with Andrew Caplin), Current Opinion in Neurobiology, August 2008, 18(2): 197-202

Underlying Pillars of Utility' which tried to come up with general mappings from real world properties to utility. This literature didn't get very far.

2. The process of deriving the axiomatic representation of a model gives you a complete list of the implications of that model. It therefore makes it very clear what behavior is and is not in line with the model. This makes it very clear how to design tests of your model, and how the implication of your model relates to those of other models.

Note, however, that this doesn't mean that I think that one should *only* use axiomatic methods to test models, just that they do provide a useful additional tool. Without performing this step, it is very easy to get confused about what you are really saying when you write down a new model, as the following example[3] illustrates.

Example 1 *There has long been evidence that people's behavior is reference dependent, in the sense that what people choose is affected by what their reference point is (the classic example of this is the endowment effect - if people are given a mug and asked if they want to exchange it for a chocolate bar then most will keep the mug. If people are given the chocolate bar and asked if they want to exchange it for the mug, most will stick with the chocolate bar). Generally, when we try to model reference dependent preferences, we assume that we know what the DM's reference point is in any given situation. However, this is a strong assumption, and a recent paper[4] took a different approach. It assumes that people had utility functions of the form* $U : X \times X \to \mathbb{R}$*, where* $U(x, z)$ *is the utility of choosing alternative* x *when* z *is the status quo. They further assume that people will choose from a set those objects which form a* ***personal equilibrium****. That is, the objects such that, if that object is the status quo, then it is the preferred object in the choice set. In other words*

$$C(A) = \{x \in A | U(x, x) \geq U(y, x) \ \forall \ y \in A\}$$

We will call this the general personal equilibrium (PE) model. The paper also provides a specific version of the model, which adds two assumptions:

1. *U has the following functional form:*

$$U(x, y) = \sum_{k \in K} u_k(x) + \sum_{j \in K} \mu(u_j(x) - u_j(y))$$

[3] This example is stolen from "The Case for Mindless Economics" by Gul and Pesendorfer

[4] "A Model of Reference Dependent Preferences" By Koszegi and Rabin - 2005

where K indexes the hedonic dimensions of the various objects and μ is an increasing function with $\mu(0) = 0$

2. *'Status quo bias'*

$$\begin{aligned} U(x,y) &\geq U(y,y) \\ &\Rightarrow U(x,x) > U(y,x) \end{aligned}$$

This condition states that if x is at least as good as y when y is the status quo, then x must be strictly better than y when x is the status quo.

We will call a general PE model that satisfies these two assumptions a special PE model

The concept of a personal equilibrium seems like an interesting way of modelling reference dependence, and one that may be worth studying. However, from just looking at the above description of the model, it is hard to tell what the behavioral implications of the model are, and what the difference is between the special and general PE models.

Unfortunately, Gul and Pessendorfer show that (i) the general and specific PE model have the same implications, and (ii) both are equivalent to dropping the assumption of transitivity from the standard rational model - in other words, a choice data set allows for a specific (or general) PE model if and only if it can be rationalized with a binary relation that is complete (but not necessarily transitive) . This means that the PE models are not necessarily particularly interesting ***unless you have a richer data set.***

We will now state this result in the form of a proposition

Proposition 1 *Let $C : 2^X/\varnothing \to 2^X/\varnothing$ be a choice function on a finite X The following statements are equivalent*

1. *(General PE model): There exists a general PE utility function $U : X \times X \to \mathbb{R}$ such that*

$$C(A) = \{x \in A | U(x,x) \geq U(y,x) \ \forall \ y \in A\}$$

2. *There exists a complete, reflexive binary relation $\succeq$ such that*

$$C(A) = \{x \in A | x \succeq y \ \forall \ y \in A\}$$

3. (Special PE model) There exists a special PE utility function $U : X \times X \to \mathbb{R}$ such that

$$C(A) = \{x \in A | U(x,x) \geq U(y,x) \ \forall \ y \in A\}$$

Proof. *The proof comes in three parts* ■

Proof (1 implies 2). *Say that C admits a general PE model. Define $\succeq$ as $x \succeq y$ if and only if $U(x,x) \geq U(y,x)$. This must be complete, as, for any x, y, if C is a choice function then either $x \in C(\{x,y\})$ or $y \in C(\{x,y\})$, and this implies that either $U(x,x) \geq U(y,x)$ or $U(y,y) \geq U(x,y)$. Furthermore, note that, by assumption*

$$\begin{aligned} C(A) &= \{x \in A | U(x,x) \geq U(y,x) \ \forall \ y \in A\} \\ &= \{x \in A | x \succeq y \ \forall \ y \in A\} \end{aligned}$$

■

Proof (2 implies 3). *Let $n = |X|$. Let $x \succ y$ indicate the asymmetric part of $\succeq$, and let $K = X \times X$ (i.e., the set of hedonic states are indexed by the cross product of X) For $k \in (w,z) \in K$, define the utility function*

$$u_{(w,z)} : X \to \{-2, 0, 2, 3\}$$

as

$$u_{(w,z)} = \left\{ \begin{array}{c} 3 \text{ if } x = y = z \\ 2 \text{ if } x = w \text{ and } w \succ z \\ -2 \text{ if } x = z \text{ and } w \succ z \\ \text{otherwise} \end{array} \right\}$$

Define μ as follows

$$\mu(t) = \left\{ \begin{array}{c} 16nt \text{ if } t \in \{-4, -3, 4\} \\ t \text{ if } t \in \{-2, 0, 2, 3\} \end{array} \right\}$$

You should check that $\mu(t)$ satisfies the necessary properties.

Let

$$U(x,y) = \sum_{k \in K} u_k(x) + \sum_{j \in K} \mu(u_k(x) - u_k(y))$$

Next, we need to show that $U(x,x) \geq U(y,x)$ iff $x \succeq y$. To see this, first note that

$$2n \geq \sum_{k \neq (x,x)} u_k(x) \geq -2n$$

And second note that, if we define $K_{x,y} = K/\{(y,y),(x,x),(x,y),(y,x)\}$ *and note that for* $k \in K_{x,y}$

$$2 \geq u_k(x) - u_k(y) \geq -2$$

Thus, it must be the case that

$$\begin{aligned} & 4n \\ \geq & \sum_{K_{x,y}} u_k(x) - u_k(y) \\ = & \sum_{K_{x,y}} \mu(u_k(x) - u_k(y)) \\ \geq & -4n \end{aligned}$$

Now assume that $x \succeq y$. *This implies that* $\mu(u_{x,y}(y) - u_{x,y}(x)) \leq 0$ *and* $\mu(u_{x,y}(y) - u_{x,y}(x)) \leq 0$. *Thus,*

$$\begin{aligned} & U(x,x) - U(y,x) \\ = & \sum_{k \in K} (u_k(x) - u_k(y)) - \sum_{k \in K} \mu(u_k(x) - u_k(y)) \\ \geq & -4n - \sum_{K_{x,y}} \mu(u_k(x) - u_k(y) \\ & -\mu\left(u_{\{x,x\}}(x) - u_{\{x,x\}}(y)\right) \\ & -\mu\left(u_{\{y,y\}}(x) - u_{\{y,y\}}(y)\right) \\ \geq & -8n + 48n - 3 > 0 \end{aligned}$$

Now, if $y \succ x$, *then* $\mu(u_{x,y}(y) - u_{x,y}(x)) = 0$ *and* $\mu(u_{x,y}(y) - u_{x,y}(x)) = 64n$, *so the above becomes*

$$\begin{aligned} & U(x,x) - U(y,x) \\ = & \sum_{k \in K} (u_k(x) - u_k(y)) - \sum_{k \in K} \mu(u_k(x) - u_k(y)) \\ \leq & 4n - \sum_{K_{x,y}} \mu(u_k(x) - u_k(y) \\ & -\mu\left(u_{\{x,x\}}(x) - u_{\{x,x\}}(y)\right) \\ & -\mu\left(u_{\{y,y\}}(x) - u_{\{y,y\}}(y)\right) \\ & -\mu(u_{x,y}(y) - u_{x,y}(x)) \\ \leq & 8n + 48n - 3 - 64n < 0 \end{aligned}$$

So we have the representation. The final thing that we need to check is that $U(x,y)-U(y,y) \geq 0$ *implies that* $U(x,x)-U(y,x)>0$. *This follows from*

$$\begin{aligned} & U(x,x)-U(y,x) \\ \geq & U(x,x)-U(y,x)-U(x,y)+u(y,y) \\ = & -\sum_{k\in K}\mu(u_k(y)-u_k(x))+\mu(u_k(x)-u_k(y)) \\ = & -2\left(\mu(-3)+\mu(3)\right)=2(48n-3)>0 \end{aligned}$$

■

Proof (3 implies 1). *This is immediate* ■

In summary, representation theorems provide an important link between the models that we have in our head (which often include latent, or unobservable variables) and the data on which we will test these models. They tell us precisely what the testable implications of our models are for any given data set, allowing us to understand whether a model is testable on a particular data set, and whether two models do in fact have different implications. Thus, in my opinion, they play a crucial role in the interaction between economic theory and data.

It is only fair to mention that there are issues in using axioms to derive testable implications of our models. Axioms provide a very stark test: either the axioms hold, in which case the model can explain the data, or they do not, and so it cannot. Thus, one mistaken choice, or one incorrectly recorded outcome is enough to discard the entire model. Given that any actually data that we use will invariably include many instances of both, we will presumably be in a situation in which we will have to reject all feasible models that we come up with. While this is a serious issue, it is not an insurmountable one: people are currently developing techniques to determine whether axioms are 'approximately' correct, which we will discuss in a later class.