MINIMAL EXPERIMENTS†

PAUL J. HEALY* & GREG LEO**

ABSTRACT. Given a parameterized model of preferences, what choice data or experiment would identify an agent's parameter value in that model? Similarly, what data would be sufficient for testing definitely whether an agent is consistent with the model? We identify a method for finding experiments that will either classify or test a given model. We do so using a novel graph-theoretic construction: The labeled permutohedron. We then provide an algorithm that finds the "smallest" such experiment for any model.

Keywords: Experimental Design; Model Testing; Incentive Compatibility

JEL Classification: C90, C91, D01.

August 28, 2022

†The authors thank Yaron Azrieli, Gary Charness, Jennifer Pate, and attendees of the ESA Boston 2022, ESA Tucson 2021, and SEA Houston 2021 meetings for their valuable comments and feedback.

*Dept. of Economics, The Ohio State University; healy.52@osu.edu.

**Dept. of Economics, Vanderbilt University; g.leo@vanderbilt.edu.

I. Introduction

Imagine a researcher that observes agents' choices and wishes to classify them according to a given model. For example, an experimenter may want to identify each subject's risk aversion parameter under the assumption of constant relative risk aversion (CRRA). The experimenter offers the subject a variety of menus of lotteries and identifies their risk parameter based on the agent's observed choices.

In other cases, the researcher wishes to test the model, for example by looking to see whether subjects' choices are consistent with maximizing a CRRA preference. Here, the exact risk aversion parameter of the subjects are not of interest; the experimenter only cares whether the CRRA model is an accurate description of their choices.

In theory, both of these goals can be accomplished by observing agents' entire preference ordering, for example by observing choices over all possible pairs of objects. Once a subject's entire ranking over all lotteries is known, the researcher can either pin down their CRRA parameter or say whether or not their preferences are inconsistent with the CRRA model.

But field data are rarely rich enough to allow such precise inference, and in the laboratory asking subjects to make that many choices would be prohibitively time-consuming. And, for most models, learning the entire preference ordering is unnecessary; agents can often be classified and models can be tested with far less information.

In this paper, we ask, for any given model, how to identify the "minimal" amount of choice data needed either to classify agents within that model or to test the model. We do so by characterizing the set of experiments that will either classify or test a given model, and then the researcher can choose the minimal experiment from this set. The exact definition of "minimal" need not be specified here; once we identify the set of possible experiments, any (partial) ordering can be used to identify the minimal experiment.[1]

Our main results are characterizations of experiments (or, more generally, choice datasets) that successfully classify agents within a model, test the model, or both. These characterizations involve a novel graph-theoretic construction: the labeled permutohedron. This provides insights into how choice data relates to the identification and testing of models. More practically, the characterizations can be used to construct a simple algorithm that quickly finds an experiment that is minimal for classifying or testing a given model. In the sequel, we provide such an algorithm for one possible definition of minimality.

In the next section, we demonstrate the framework and the key results of our paper through several simple examples. Most of the intuition behind our characterizations is present in these examples. In Sections III–V we provide our formal framework, which extends that of Azrieli et al. (2021), and state our main characterizations. In Section VI we

[1]Some examples of "minimal" are the experiment with the fewest number of choice menus given to the subject, the smallest expected cost to the experimenter, or the minimal amount of privacy invasion incurred by subjects.

extend our results further by showing how they apply to experiments where subjects can choose more than one option from a given menu. For example, subjects may be asked to pick their top k items from each menu, rather than a single choice item.[2] In Section VII we explore additional properties of the permutohedron that might be useful in future work. In Section VIII we provide a simple algorithm for finding an experiment that is minimal in terms of the number and size of menus. Section IX concludes with a discussion.

II. Illustrative Examples

In an early economic experiment, Rousseas and Hart (1951) asked subjects to rank three plates of eggs and bacon. To construct indifference curves from their data, the authors made several assumptions about preferences, including monotonicity and convexity. We can think of each assumption as a separate *model* of preferences. In this section we demonstrate how our methods can be used to test these two models, as well as a separate model of coarse beliefs about an unknown event.

Model 1: Monotonic Preferences

In the eggs-and-bacon example, each plate can be written as an ordered pair, with the first entry giving the number of eggs and the second entry the number of pieces of bacon. Suppose the available options are $a = (3,3)$, $b = (1,2)$, and $c = (2,1)$, and the researcher is interested in testing monotonicity. This assumption requires $a \succ b$ and $a \succ c$. The (strict) rank orderings consistent with monotonicity are abc (meaning $a \succ b \succ c$) and acb (meaning $a \succ c \succ b$), while the rankings bac, bca, cab, and cba are inconsistent with monotonicity. We can therefore view monotonicity as a model M in which $M = \{abc, acb\}$ are the preferences allowable within the model, and the complementary set $M_0 = \{bac, bca, cab, cba\}$ contains the preferences outside the model.

What experiment could be used to test whether this model is true or not? In other words, how can we distinguish whether a subject's preferences are in $\{abc, acb\}$ or not? The simplest way (in terms of number of questions) is to offer the subject a menu of all three plates $\{a, b, c\}$ and ask them to choose one. Formally, the subject is given a single decision problem $D_1 = \{a, b, c\}$ and chooses their most-preferred item from that menu. If the subject chooses a then the model is validated, otherwise, it fails.

Suppose our notion of "minimal" is to use the fewest decision problems as possible, with the size of the menus being a (partial) tie-breaking rule.[3] This experiment is clearly minimal

[2] This can be incentivized by paying each chosen item with probability $1/k$; see Azrieli et al. (2020) for details.

[3] We refer to this as the *lexicographic size ordering* and formalize it in Section VIII.

under that ordering. Of course, more complex experiments could also test this model: Offering every binary menu ($D_1 = \{a,b\}$, $D_2 = \{a,c\}$, $D_3 = \{b,c\}$) would completely identify the subject's ordering, and therefore would be sufficient to test the model, but with two more decisions than is necessary. Thus, it is not minimal.

Model 2: Convex Preferences

As a second example, consider the model of (strictly) convex preferences. Suppose now the plates available are $a = (2,2)$, $b = (3,1)$, and $c = (1,3)$. Since plate a is a convex combination of the other plates, convexity requires a to be preferred to the least-preferred of plates b and c. That is, a cannot be ranked last. The set of rankings meeting this condition is $M = \{abc, acb, bac, cab\}$ and the complementary set of rankings outside the model is $M_0 = \{bca, cba\}$.

This model cannot be tested by the choice of a favorite plate from a single menu. Instead, the minimal experiment uses two menus: $D_1 = \{a,b\}$ and $D_2 = \{a,c\}$. If the subject chooses a in at least one decision, the model is validated. Otherwise, it fails.[4]

A model M may further partition the preferences into "types," or "parameters." For example, suppose the researcher is also interested in splitting the convex preferences into those that most prefer a, those that most prefer b, and those that most prefer c. We formalize this by writing model M as a partition $M = \{t_1, t_2, t_3\}$, where $t_1 = \{abc, acb\}$ is the type that most prefers a, $t_2 = \{bac\}$ is the type that most prefers b, and $t_3 = \{cab\}$ is the type that most prefers c. Again, $M_0 = \{bca, cba\}$ are the preferences outside the model.

Interestingly, it is possible to classify subjects into these types using the same experiment described above: $D_1 = \{a,b\}$ and $D_2 = \{a,c\}$. Agents of type t_1 will pick (a,a) (meaning a from D_1 and a from D_2), agents of type t_2 will pick (b,a), and agents of type t_3 will pick (a,c). Any other choice reveals M_0. Thus, this experiment both tests and classifies the model.

Types may also be associated with parameter values, or ranges of parameter values of a utility function. For instance, the utility function $u(x_1,x_2) = x_1^{\alpha} x_2^{1-\alpha}$ refines the convexity model discussed above, splitting the set $\{abc, acb, bac, cab\}$ into singleton types $t_1 = \{bac\}$, $t_2 = \{abc\}$, $t_3 = \{acb\}$, and $t_4 = \{cab\}$, associated with the parameter values $\alpha > 0.63$, $\alpha \in (0.5, 0.63)$, $\alpha \in (0.37, 0.5)$, and $\alpha < 0.37$, respectively. The minimal experiment to classify subjects into these types (assuming preferences in M_0 are impossible) is $D_1 = \{a,b,c\}$ and $D_2 = \{b,c\}$. Here, D_1 reveals the favorite element and, if it's a, then D_2 is used to identify the second-most preferred.[5]

[4]Notice that this model can be tested with a single menu if subjects are incentivized to reveal their top *two* options from $\{a,b,c\}$ (or, equivalently, to eliminate their least favorite option). We extend our results to menus of this type in Section VI.

[5]Suppose we want to classify *and* test this model. Recall that testing requires $D_1 = \{a,b\}$ and $D_2 = \{a,c\}$, while classifying requires $D_3 = \{b,c\}$. Thus, it is necessary to elicit the entire ordering in this case.

We relate experiments to models through the notion of separation. We say an experiment *separates* two rankings if subjects with those rankings make different choices in the experiment. The rankings an experiment needs to separate depends on which goal the experimenter is pursuing. Testing a model requires separating all rankings inside the model (M) from those outside the model (M_0). Classifying a model requires separating all rankings of each type ($t_i \in M$) from all rankings of the other types ($t_j \in M$). Classifying does not require separating rankings in the model from rankings outside the model, and testing does not require separating the various types inside the model.

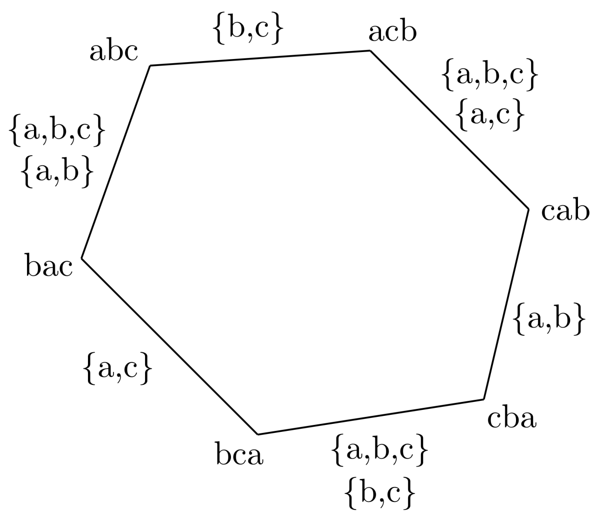

FIGURE I. The labeled permutohedron for three objects.

To understand which rankings are separated by a given experiment, we first visualize all possible rankings on a graph called the permutohedron. The permutohedron is constructed by placing each preference ranking on a vertex and connecting rankings that differ only by a single swap of adjacent pairs in the ordering. We call such rankings "neighbors." For instance, abc and acb are neighbors because they differ only in their ranking of b and c.

Next, we augment the permutohedron by labeling each edge with those menus from which the neighboring rankings would choose differently. For instance, abc and acb choose differently only from the set $\{b,c\}$, so we label the edge between abc and acb with the set $\{b,c\}$. The rankings acb and cab choose differently from both $\{a,c\}$ and $\{a,b,c\}$, so both appear on the edge between acb and cab. The labeled permutohedron for three objects is shown in Figure I.

The key results of our paper show how the labeled permutohedron can be used to characterize the experiments that test and classify any model. This is true even though the permutohedron has no direct information about what sets separate the nonadjacent rankings. Specifically, our main theorem shows that to test or classify a model, an experiment must contain at least one set from the edge between every "boundary pair" of rankings. Boundary pairs are pairs of adjacent rankings that lie in different sets in the model.

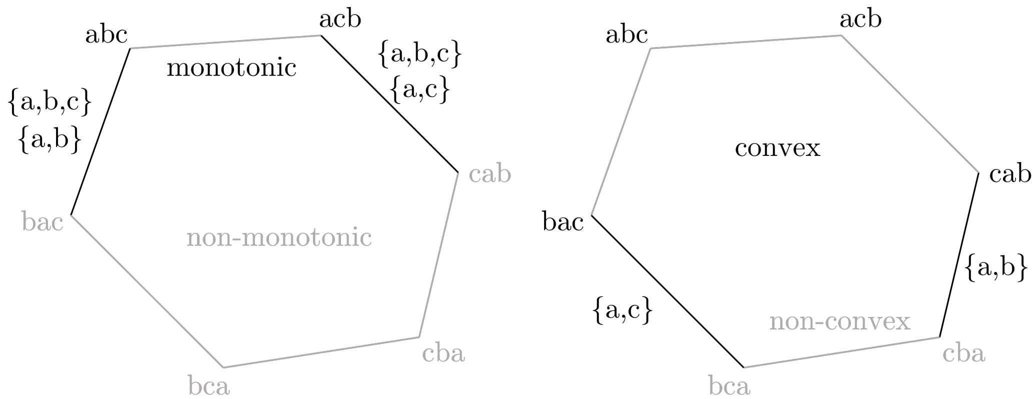

FIGURE II. Boundary pairs in the monotonicity model (left panel) and convexity model (right panel).

To illustrate, Figure II highlights the boundary pairs for the monotonicity and convexity examples discussed above. The edges between boundary pairs are shown in bold. In the case of monotonicity, the boundary pairs are $\{abc, bac\}$ (because $abc \in M$ and $bac \in M_0$) and $\{acb, cab\}$ (because $acb \in M$ and $cab \in M_0$). An experiment will test this model if and only if it contains a menu from each of the two edges connecting these boundary pairs. For example, $D_1 = \{a, b\}$ and $D_2 = \{a, c\}$ will test this model. But since $\{a, b, c\}$ appears on both boundary pair edges, the experiment $D_1 = \{a, b, c\}$ also tests the model. Since this experiment uses the fewest number of menus, it is minimal in that sense.[6]

Moving to convexity, there are again two boundary pairs: $\{bac, bca\}$ and $\{cab, cba\}$. Since the first edge contains only the set $\{a, c\}$, it must be included in the experiment. The second edge contains only the set $\{a, b\}$, so it also must be included in the experiment. Thus, every experiment that tests this model must include these two sets. Clearly, the minimal experiment involves just these two.

Testing a model M is equivalent to classifying subjects into one of two types: those in the model ($t_1 = M$), and those outside the model ($t_2 = M_0$). An experiment will classify subjects into t_1 and t_2 if and only if it contains sets from each boundary pair's edge, as in Figure II. Classifying a model with more types works similarly: for every pair of types, t_i and t_j, find all boundary pairs between them. An experiment that classifies subjects in this model must contain at least one menu from all such boundary pairs.

Additional complications arise when the model being classified does not include all possible preferences. As an extreme example, suppose $t_1 = \{abc\}$, $t_2 = \{cba, cab\}$, and all other

[6]Experiments $D_1 = \{a, b, c\}$, $D_2 = \{a, c\}$ and $D_1 = \{a, b, c\}$, $D_2 = \{a, b\}$ would also test the monotonicity model, as would any experiment that contains additional sets. But none of these would be minimal.

rankings are outside the model. Here, there are no boundary pairs between t_1 and t_2. In Section V we demonstrate how to modify the labeled permutohedron in such cases. We first identify the shortest path between t_1 and t_2, which would go from abc to cab via acb. We then construct a new graph that connects abc directly to cab, since they were connected indirectly via this shortest path. The sets labeled on the new edge between them are all the sets that were labeled on that shortest path, which would be $\{b,c\}$, $\{a,b,c\}$, and $\{a,c\}$. Rankings outside the model (and their corresponding edges) are then deleted. In Section V we show that an experiment classifies a model if and only if it contains sets from every boundary pair on this "restricted" graph.

In Section VI we extend our theorems to include choice tasks where subjects are asked to select their top k_i favorite objects from each set D_i. For instance, the minimal experiment for testing the convexity model above requires subjects to choose one plate from two sets. However, if we extend the possible experiment with these choose-k menus, it can be tested with a single choice task in which subjects choose their favorite two plates from $\{a,b,c\}$.

While these examples are simple, the logic generalizes to any model over a finite set of alternatives X. In the next section, we demonstrate how our framework and theorems can be applied in a more interesting setting, generating novel methods for belief elicitation.

Model 3: Ranges of Beliefs

Suppose a researcher wants to elicit a subjective belief about the probability an event E will occur. The researcher is interested in categorizing the belief p into three categories: $p \in [0,0.4)$, $p \in (0.4,0.6)$, and $p \in (0.6,1]$. Suppose there are three relevant choice objects available: $l_{0.6}$ pays \$10 with objective probability 0.6, t pays \$10 if event E is true, and f pays \$10 if E is false.

The three belief categories can be represented by a model with three types: Beliefs $p \in [0,0.4)$ correspond to the singleton type $t_1 = \{f l_{0.6} t\}$ (meaning $f \succ l_{0.6} \succ t$), beliefs $p \in (0.4,0.6)$ correspond to the type $t_2 = \{l_{0.6} f t, l_{0.6} t f\}$, and beliefs $p \in (0.6,1]$ correspond to the type $t_3 = \{t l_{0.6} f\}$. The rankings outside the model are those for which $l_{0.6}$ is ranked last ($M_0 = \{t f l_{0.6}, f t l_{0.6}\}$).

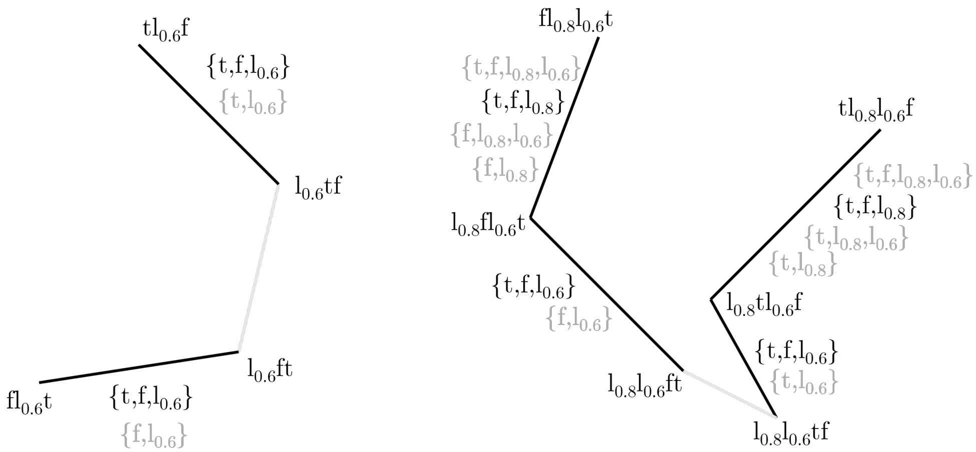

FIGURE III. The restricted permutohedra for three-category (left) and five-category (right) belief elicitation. Only the edges between boundary pairs (shown in bold) have been labeled. Sets used in the relevant minimal experiment are shown in bold.

Classifying subjects in this model means the researcher assumes preferences in M_0 are impossible.[7] Thus, we apply the restricted permutohedron described above. For the purposes of this example, the restricted permutohedron is simply the graph induced by removing the vertices in M_0 from the permutohedron.[8] The restricted permutohedron for this model is shown in the left panel of Figure III.

The minimal experiment for the three-category belief elicitation involves just one set: $D_1 = \{t, f, l_{0.6}\}$. This set appears on both of the edges between the two boundary pairs on the restricted permutohedron. The experiment might appear this way:

Choose how you would most like to be paid. At the end of the experiment, you will receive your chosen payment option.

\$10 if E occurs	\$10 if E does not occur	\$10 with a 60% chance

Suppose we expand the categorization to have five types: $p \in [0, 0.2)$, $p \in (0.2, 0.4)$, $p \in (0.4, 0.6)$, $p \in (0.6, 0.8)$, and $p \in (0.8, 1]$. Letting $l_{0.8}$ be the lottery that pays \$10 with probability 0.80, this can be represented by the following model:[9]

$$t_1 = \{tl_{0.8}l_{0.6}f\}, t_2 = \{l_{0.8}tl_{0.6}f\}, t_3 = \{l_{0.8}l_{0.6}tf, l_{0.8}l_{0.6}ft\}, t_4 = \{l_{0.8}fl_{0.6}t\}, t_5 = \{fl_{0.8}l_{0.6}t\}$$

[7]For example, these rankings are impossible under the assumption that each individual has a subjective probability p about the likelihood of event E and orders all bets by their probability of \$10.

[8]We note that, while the construction of the restricted permutohedra in these two examples involves simply removing vertices (and relevant edges) from the full permutohedron, there are models in which new edges must be created as well. The details of this are provided in Section V

[9]The rankings outside the model M_0 have not been written, but are the other 18 rankings. They are the 12 rankings with $l_{0.6} \succ l_{0.8}$ and the 6 rankings with $l_{0.6}$ ranked last.

The restricted permutohedron for this model, shown in the right panel of Figure III, is the four-object permutohedron with the 18 rankings from M_0 removed.

From this, we can see that the minimal experiment is $D_1 = \{t, f, l_{0.6}\}$ and $D_2 = \{t, f, l_{0.8}\}$.[10] The experiment might appear this way:

In each row below, choose how you would most like to be paid. At the end of the experiment, one row will be chosen at random, and you will receive your chosen payment option.

\$10 if E occurs	\$10 if E does not occur	\$10 with an 80% chance
\$10 if E occurs	\$10 if E does not occur	\$10 with a 60% chance

It is possible to find the minimal experiment for belief elicitation with a larger number of categories as well. This can be done analytically or computationally using the algorithm provided in Section VIII. As long as there is an odd number of categories and those categories are symmetric around 0.5 (as they are in these two examples), the minimal experiment has a similar structure. Specifically, each menu offers three options: bet t, bet f, and some l_p. We call these *ternary price lists*. The more rows offered in a ternary price list, the more refined is the model of beliefs that can be classified. Exact probabilities (down to 1% precision) can be elicited using 50 rows, with the rightmost option ranging from "\$10 with a 100% chance" to "\$10 with a 50% chance." For comparison, the usual binary price list (comparing only "\$10 if E occurs" to "\$10 with a p% chance") would require a full 100 rows. Adding the middle option cuts the number of rows needed in half.

Ternary price lists are simple, and minimal for these types of models. However, there are other ways of eliciting probabilistic beliefs. The most popular in experimental economics is the binarized quadratic scoring rule (Savage, 1971; Hossain and Okui, 2013). This procedure asks a subject their belief p and maps this into a compound lottery $L_{1-(1-p)^2}$ that pays \$10 (or, any fixed prize) with probability $1-(1-p)^2$ if the event occurs, and \$10 with probability $1-p^2$ if it does not.

It is possible to analyze this procedure through the scope of our framework. Denote the set of possible compound lotteries in the BQSR as $\mathscr{L} = \{L_{1-(1-p)^2} : p \in [0,1]\}$. The procedure consists of a single choice from the (infinite) set $\mathscr{L}$. If subjects reduce compound lotteries then truth-telling is optimal. And this experiment is minimal for a model in which agents reduce compound lotteries since each belief corresponds to a unique favorite lottery, which is the truth-telling lottery.

The BQSR involves a single choice from a large set of compound lotteries. A ternary price list involves several rows of choices over three simple lotteries. Formally, they cannot

[10] Testing and classifying this model can be achieved with four sets: $D_1 = \{l_{0.8}, f, t\}$, $D_2 = \{l_{0.8}, l_{0.6}\}$, $D_3 = \{l_{0.6}, f\}$, and $D_4 = \{l_{0.6}, t\}$.

be compared, since one uses compound lotteries and one uses simple lotteries. Both are minimal for their respective models. Which one a researcher prefers depends on whether compound lotteries are available, whether they believe subjects reduce compound lotteries, and whether they prefer asking fewer questions or having questions with fewer alternatives.

In the next section we provide our general framework, which encompasses all of these examples.

III. The Framework

Given is a finite set X of $m \geq 2$ alternatives, with typical elements denoted by a, b, c, and so on. The set of all complete strict orderings of X (the orderings that are complete, reflexive, transitive, and antisymmetric) is given by $\mathscr{P}$. A typical element of $\mathscr{P}$ is denoted by P.[11] To economize notation, we use abc to denote the P such that aPb and bPc, for example.

A *model* $M = (t_1, \ldots, t_n, M_0)$ is a partition of $\mathscr{P}$, where each $t_i \subseteq \mathscr{P}$ ($t_i \neq \emptyset$) is referred to as a *type* within the model and $M_0 \subseteq \mathscr{P}$ is the set of orders not included in model M. When $P \in M_0$ the interpretation is that model M assumes no subject could have ordering P. For example, if X is a set of simple lotteries and M is the expected utility model then each t_i identifies a unique ordering with parallel, linear indifference curves on the simplex and M_0 contains all non-expected-utility orderings. Abusing notation, write $P \in M$ if $P \notin M_0$, in which case we say that P is included in model M. We say a model is *complete* if $M_0 = \emptyset$, and *restricted* otherwise. When $P \in M$ let $t(P)$ be the type containing P; set $t(P) = M_0$ if $P \in M_0$.

An *experiment* is a family of sets $\mathscr{D} = \{D_1, \ldots, D_n\}$ such that $D_i \subseteq X$ and $D_i \neq D_j$ for all i and $j \neq i$. The interpretation is that each D_i is a menu from which the subject must choose their most-preferred element. We define the following choice function:

$$\mathrm{dom}_P(X') = \{x \in X' : (\forall y \in X')\ xPy\}.$$

Since all orders are assumed to be antisymmetric, $\mathrm{dom}_P(X')$ will always contain a single element.

We now define how a model distinguishes between two orders, and compare that to how an experiment distinguishes between those orders.

Definition 1 (*Differentiated Pair*)**.** Fix a model $M = (t_1, \ldots, t_n, M_0)$. Two orders P and P' are *differentiated by M* (or, $\{P, P'\}$ is a *differentiated pair*) if $t(P) \neq t(P')$.

Definition 2 (*Separated Pair*)**.** Fix an experiment $\mathscr{D}$. Two orders P and P' are *separated by $\mathscr{D}$* (or, $\{P, P'\}$ is a *separated pair*) if there exists some $D_i \in \mathscr{D}$ such that $\mathrm{dom}_P(D_i) \neq \mathrm{dom}_{P'}(D_i)$.[12]

[11] To be clear, these are strict rankings with the added requirement that every alternative is comparable to itself. Thus, aPb and bPa implies $a = b$.

[12] Note that Definitions 1 and 2 apply to any pair P and P', including those for which $P \in M$ and $P' \in M_0$.

Roughly, our goal is to find an experiment whose separated pairs are a superset of the differentiated pairs of the model. Formally, every experiment $\mathscr{D}$ defines a partition $R_{\mathscr{D}} = (r_1,\ldots,r_k)$ of $\mathscr{P}$ such that P and P' are in the same partition element if and only if they are not separated by $\mathscr{D}$. Letting $r(P)$ be the partition element that contains P, the partition is formally defined by: $P \in r(P')$ if and only if for every $D_i \in \mathscr{D}$ we have $\text{dom}_P(D_i) = \text{dom}_{P'}(D_i)$. We refer to this as the *experiment partition* for experiment $\mathscr{D}$.

We can now give our main definitions of classifying and testing models using an experiment. The idea is that the experiment partition should be a refinement of the relevant partition given by the model, which is $\{M, M_0\}$ when testing and $\{t_1,\ldots,t_n\}$ when classifying.

Definition 3 (*Classifying*)**.** An experiment $\mathscr{D}$ *classifies agents according to model M* (or, more simply, *classifies M*) if every $P \in M$ and $P' \in M$ that are differentiated by M are separated by $\mathscr{D}$.

In other words, if P and P' belong to different types in the model (but not M_0) then there is some $D_i \in \mathscr{D}$ for which they will choose differently. Thus, the experimenter can use an agent's choices to identify their type.

Definition 4 (*Testing*)**.** An experiment $\mathscr{D}$ *tests model M* if all $P \in M$ and $P' \in M_0$ are separated by $\mathscr{D}$.

In words, testing a model simply means that the agent's choices inform the experimenter whether their preference P is included in the model or belongs to M_0.

An important difference between testing and classifying is that when classifying, we only consider orders P and P' that are both in M. It is as though the researcher assumes that any $P \in M_0$ will not be observed and is only interested in the subject's type t_i. When testing, the experimenter is only interested in learning whether $P \in M$, and not interested in learning the agent's type. An experiment *tests and classifies* a model if it accomplishes both.

Testing a model can equivalently be viewed as classifying the subject into one of two types: those consistent with the model, and those not. Formally, testing model $M = (t_1,\ldots,t_n,M_0)$ is equivalent to classifying the complete model $M' = (t'_1,t'_2)$ defined by $t'_1 = \bigcup_i t_i$ (those consistent with M) and $t'_2 = M_0$ (those not consistent with M). Thus, theoretical conditions for testing a model are very similar to those needed for classifying a complete model. Classifying a restricted model, however, is fundamentally different, so its results are presented separately.

An *experiment ordering* $>$ is a strict partial order on the set of experiments. When $\mathscr{D}' > \mathscr{D}$ we say that $\mathscr{D}$ is smaller than $\mathscr{D}'$. An experiment $\mathscr{D}$ is *minimal for testing M* if $\mathscr{D}$ tests M and there is no $\mathscr{D}'$ that tests M such that $\mathscr{D} > \mathscr{D}'$. Analogously, $\mathscr{D}$ is *minimal for classifying M* if it classifies M and no smaller experiment classifies M.

The Premutohedron

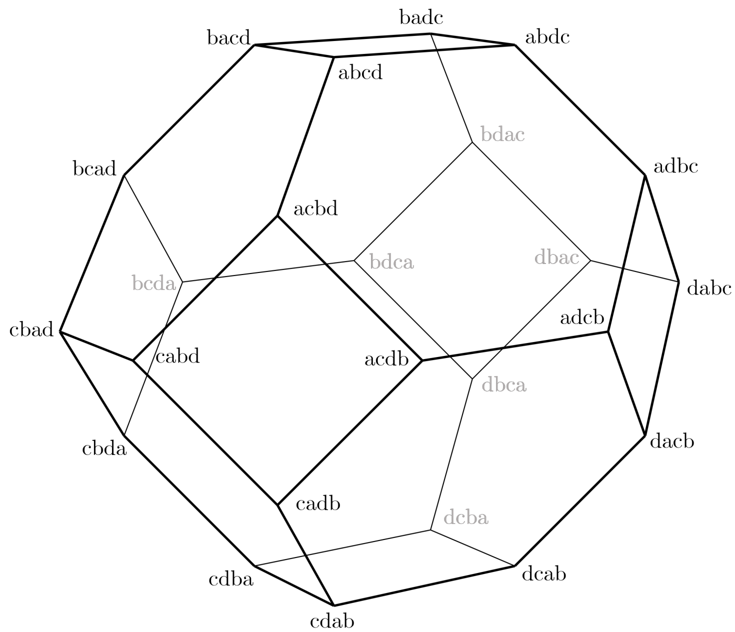

FIGURE IV. The permutohedron for four objects $X = \{a, b, c, d\}$

We now introduce the geometric structure we use to characterize experiments that test and classify models.

The set of transpositions between two orderings P and P' is given by

$$T(P,P') = \left\{\{x,x'\} \subseteq X : \mathrm{dom}_P(\{x,x'\}) \neq \mathrm{dom}_{P'}(\{x,x'\})\right\}.$$

We say P and P' are *neighbors* if $|T(P,P')| = 1$.

The *transposition graph* is a tuple $(\mathscr{P},\mathscr{E})$ in which all orderings in $\mathscr{P}$ are nodes and all edges in $\mathscr{E}$ connect two neighbors: $\mathscr{E} = \{\{P,P'\} : |T(P,P')| = 1\}$. This graph can be represented as a polytope in $|X|$-dimensional Euclidean space by mapping each ranking into a vertex with coordinates given by the position of the relevant object in the ranking. For instance, if $abcd$ is mapped to $(1,2,3,4)$ then $cabd$ is mapped into $(2,3,1,4)$. The resulting polytope is known as the *permutohedron*.[13] Since the sum of the coordinates is fixed for any ranking, the permutohedron lies completely in a $|X|-1$ dimensional simplex.[14]

The *labeled permutohedron* is a tuple $(\mathscr{P},\mathscr{E},L)$, which consists of a graph with nodes $\mathscr{P}$ and edges $\mathscr{E}$ as described above, but with edge labels $L : \mathscr{E} \twoheadrightarrow 2^X$ defined as follows: For any

[13]Berge (1971) attributes this name to Guilbaud and Rosenstiehl (1963).

[14]To simplify understanding in our context, we label the vertices with their associated rankings, rather than vertex coordinates as is common elsewhere. When the vertices are associated with permutations of the objects X, the graph is the Cayley graph of the symmetric group $S_{|X|}$ generated by the $|X|-1$ possible adjacent transpositions. Since the polytope and the Cayley graph are isomorphic, "permutohedron" is often used to refer to both objects. For instance, our usage is consistent with Berge (1971).

edge $E = \{P, P'\} \in \mathscr{E}$, $L(E) = \{S \subseteq X : \mathrm{dom}_P(S) \neq \mathrm{dom}_{P'}(S)\}$. That is, the edges are labeled with all the sets for which the neighboring rankings choose differently. Note that an experiment $\mathscr{D}$ separates neighbors P and P' if there exists some $D_i \in \mathscr{D}$ such that $D_i \in L(\{P, P'\})$; this will be useful in our main result.

A *path* W between P and P' is a finite sequence of nodes $(P_1 \ldots, P_n)$ with $P_i \neq P_j$ for $i \neq j$ such that $P_1 = P$, $P_n = P'$ and $\{P_i, P_{i+1}\} \in \mathscr{E}$. A path traverses n nodes and $n-1$ edges. The *length* of path W is defined as $n-1$. Let $\mathscr{E}(W)$ be the set of edges traversed by path W. A path W between P and P' is *shortest* if there is no other path between P and P' that has a smaller length. Shortest paths may not be unique.

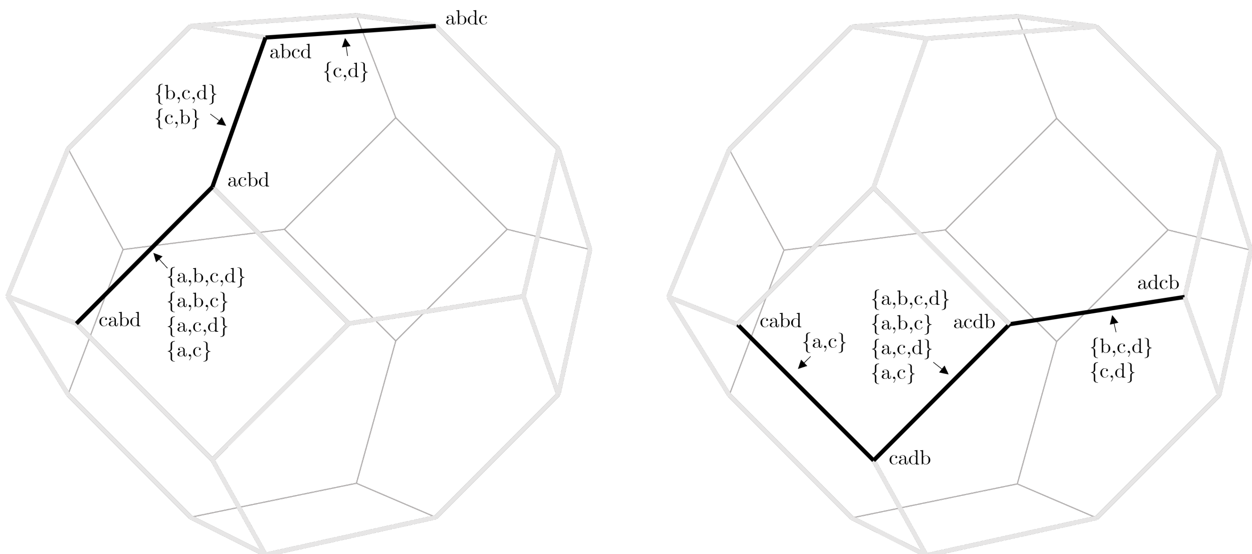

FIGURE V. The shortest path from $abcd$ to $cabd$ and one of the two shortest paths from $cabd$ to $adcb$. The edges have been labeled along each path.

Definition 5 (*Convexity*)**.** A set of rankings S is convex if, for every pair $P, P' \in S$, every shortest path from P to P' is contained in S. Additionally, we call a partition of $\mathscr{P}$ convex if every set in the partition is convex.

Experiments and Convexity

We now discuss the relationship between experiments and the geometry of the permutohedron. To help visualize this, we introduce the following definition.

Definition 6 (*Graph Induced by Experiment $\mathscr{D}$*)**.** The graph induced by experiment $\mathscr{D}$ is the labeled permutohedron with edges between rankings separated by $\mathscr{D}$ removed.

The graph induced by experiment $\mathscr{D}$ consists of distinct components, where the rankings contained in a particular component correspond exactly to some element of the experiment partition $R_{\mathscr{D}}$. In Figure VI, we show the graphs induced by four different experiments on the set $X = \{a,b,c,d\}$. This figure shows some of the complex ways that even simple experiments can partition the set of rankings.

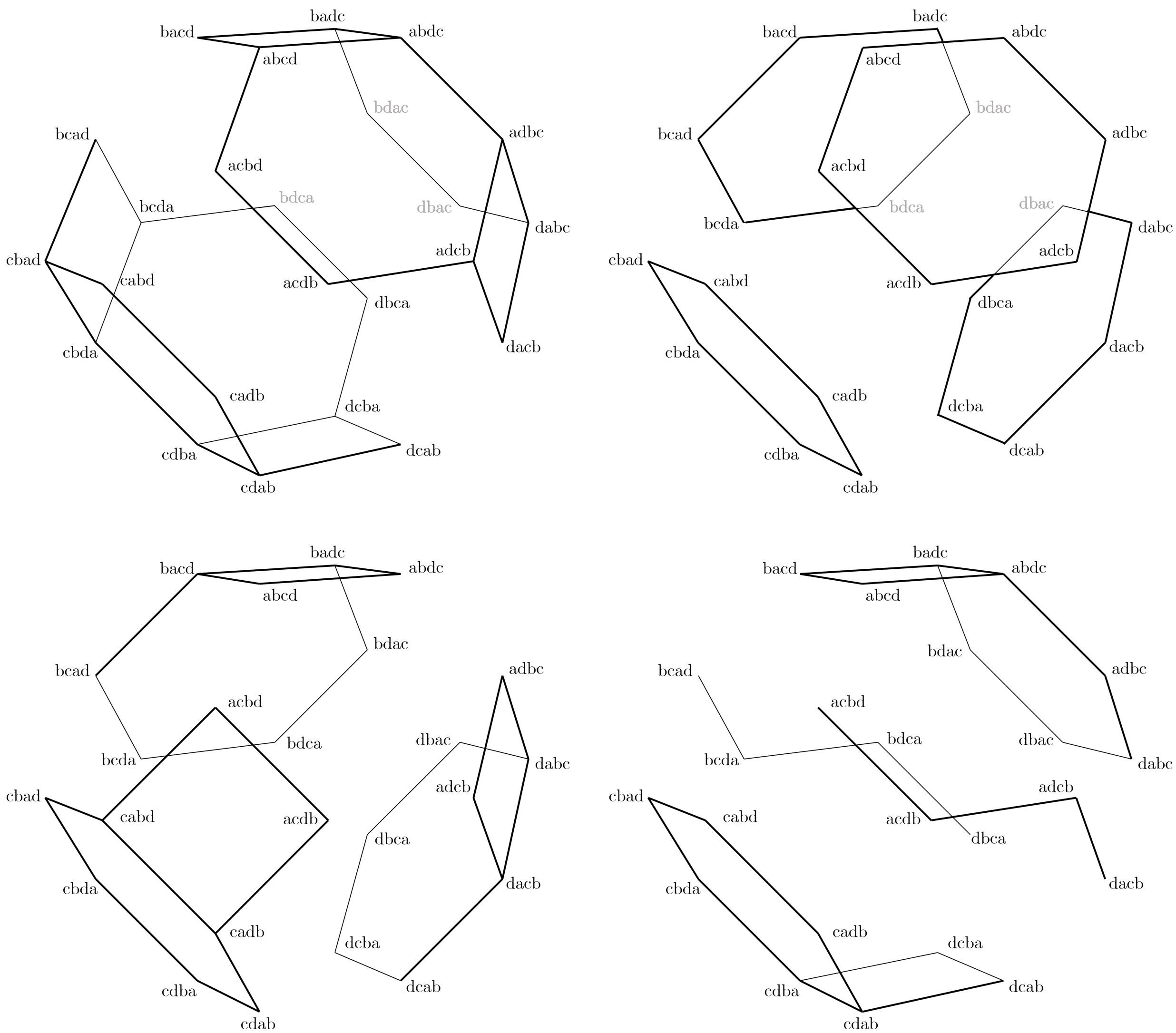

FIGURE VI. Induced graphs for experiments (clockwise) $\mathscr{D} = \{a,c\}, \mathscr{D} = \{a,b,c,d\}, \mathscr{D} = \{b,c,d\}, \mathscr{D} = \big\{\{a,c\},\{c,b\}\big\}$.

As can be seen in Figure VI, there is a lot of structure in the way that experiments partition the set of rankings. For our purposes, the most important regularity is that every experiment partition must be convex (with respect to the full permutohedron).[15] This implies that each component of the induced graph retains all the shortest paths on the full permutohedron between the rankings in that set.

Take, for example, the experiment $\mathscr{D} = \{a,b,c,d\}$ shown in the top right of Figure VI. The experiment separates every pair of rankings with a different top object, and thus partitions the rankings into the four sets defined by those top objects. This induces a graph made up of four disconnected hexagonal components, each isomorphic to the three-object permutohedron. Though it is difficult to visualize, this also provides some insight into the recursive structure of higher dimensional permutohedron. The five object permutohedron, for instance, contains five subgraphs isomorphic to the four object permutohedron shown in figure IV.

We now prove this important property of the geometry of experiments.

Proposition 1 (*Experiments are Convex*)**.** Every experiment partition $R_{\mathscr{D}}$ is convex.

Proof. The proof involves first characterizing the shortest paths between rankings via transpositions. Recall that $T(P,P')$ is the set of transpositions between P and P'.

Lemma 1 (*Adjacent Transpositions*)**.** If $T(P,P')$ is non-empty, then there must be an adjacent pair of objects in the ranking P that is transposed in P'.

Proof of Lemma 1. Assume otherwise. Let x and x' be a transposed pair in P and P'. Let $x_1, x_2, ..., x_n$ be a sequence of objects that are adjacent in the ranking P such that $x_i P x_{i+1}$ and such that $x_1 = x$ and $x_n = x'$. By assumption, Since x and x' are transposed in P' but no adjacent pair in P is transposed, we have $x_1 P' x_2 P' \ldots P' x_n P' x_1$, which contradicts the fact that each ranking must be acyclic. □

Lemma 2 (*Length of Shortest Paths*)**.** The length of any shortest path between P and P' is $|T(P,P')|$.

Proof of Lemma 2. Since P and P' differ by $|T(P,P')|$ transpositions, and each edge involves only a single transposition, the distance must be at least $|T(P,P')|$. Since each edge separates two rankings that differ only by a single transposition, that transposition must involve objects that are adjacent in each ranking. Thus, the claim is equivalent to the fact that any ranking can be transformed into any other ranking using $|T(P,P')|$ adjacent transpositions. Construct a sequence of rankings by the following procedure. Let $P_1 = P$ and for every P_i pick an adjacent pair of objects in P_i that is transposed in P'. By Proposition 1 such

[15] We note that convex partitions are not a characterization of experiments. There are convex partitions that are not induced by an experiment. In the language of Azrieli et al. (2021), such partitions are not *exactly elicitable*. This holds even for the extended experiments discussed in Section VI. While a characterization of experiments in terms of the possible partitions is outside the scope of this paper, we draw attention to the symmetries of the connected subgraphs shown in Figure VI.

a pair will always exist as long as $P_i \neq P'$, and because only adjacent swaps are made, $T(P_i,P') \subset T(P_{i+1},P')$. Thus, the sequence transforms P into P' with $|T(P,P')|$ adjacent transpositions.[16] □

Since the shortest path between P and P' has $|T(P,P')|$ edges, this is also the graph distance between P and P'. Next, we prove an important lemma about the sets of size two appearing on any shortest path between two rankings. To that end, for any path W, let $L(W)$ be the union of $L(E)$ for every edge in $\mathscr{E}(W)$.

Lemma 3 (*Shortest Paths and Adjacent Transpositions*)**.** If W is a shortest path between P and P' then every set $S \in T(P,P')$ appears exactly once in $L(W)$. Furthermore, if $S \notin T(P,P')$ and $|S| = 2$ then $S \not\in L(W)$.

Proof of Lemma 3. Every edge label contains exactly one set with $|S| = 2$ associated with the adjacent transposition between the neighboring rankings attached by that edge. If a set $S \in T(P,P')$ does not appear along W then, for every ranking $\tilde{P}$ along W, $\text{dom}_{\tilde{P}}(S)$ is the same. Thus, $\text{dom}_P(S) = \text{dom}_{P'}(S)$ which contradicts that $S \in T(P,P')$. Thus, every $S \in T(P,P')$ must appear at least once, but since the length of W is $|T(P,P')|$ by Lemma 2, and each edge had only one set on its label with $|S| = 2$, every set in $S \in T(P,P')$ must appear exactly once. □

We are now ready to prove Proposition 1 (experiments are convex). Suppose it were false. Then there is some set in $R_{\mathscr{D}}$ that is non-convex. Thus, some pair of rankings P and P' are such that $P' \in r(P)$ but there is some shortest path W between them that does not remain inside $r(P)$.

There must be some P'' on W such that $r(P'') \neq r(P)$, thus there is some set $D_i \in \mathscr{D}$ for which $x = dom_P(D_i) \neq dom_{P''}(D_i) = x''$. However, since $r(P) = r(P')$, $dom_P(D_i) = dom_{P'}(D_i) = x$. x and x' must be inverted at least twice on the path W and so the set $\{x,x''\}$ appears at least twice on some shortest path from P to P', contradicting Lemma 3.

□

IV. Classifying Complete Models & Testing Models

The Main Theorem

Recall that $\{P,P'\}$ is a differentiated pair if P and P' are assigned to different types in the model, and that the model is classified by an experiment if the experiment separates every differentiated pair. The main theorem shows that it is sufficient to check only that the experiment separates those differentiated pairs that are neighbors in the permutohedron. We call these *boundary pairs*.

[16] This algorithm is known as the *bubble sort* in the computer science literature (Astrachan, 2003).

Definition 7 (*Boundary Pairs*)**.** A pair $\{P,P'\}$ is a *boundary pair* for model M if it is a differentiated pair such that P and P' are neighbors in the permutohedron.

Theorem 1 (*Characterization of Experiments that Classify Complete M*)**.** Experiment $\mathscr{D}$ classifies a complete model $M = (t_1,\dots,t_n)$ if and only if $\mathscr{D}$ separates every boundary pair for model M.

Proof of Theorem 1. Necessity is simple: If $\mathscr{D}$ classifies M then *all* differentiated pairs are separated by $\mathscr{D}$, and so every boundary pair must also be differentiated.

For sufficiency, note that for any experiment $\mathscr{D}$ we can define the partition $R_{\mathscr{D}} = (r_1,\dots,r_k)$ of $\mathscr{P}$ such that P and P' are in the same partition element if and only if they are not separated by $\mathscr{D}$. Let $r(P)$ be the partition element containing order P.

Lemma 4 ($R_{\mathscr{D}}$ *Refines M*)**.** If $\mathscr{D}$ classifies M then $R_{\mathscr{D}}$ is a refinement of M, meaning every $r_i \in R_{\mathscr{D}}$ is a subset of some $t_i \in M$

Proof of Lemma 4. The proof of this lemma is by contradiction: If $R_{\mathscr{D}}$ were not a refinement of M then there would be an r_i that intersects two different types t_i and t_j. But then there would be some differentiated pair $P \in t_i$ and $P' \in t_j$ such that $r(P) = r(P') = r_i$, meaning $\mathscr{D}$ fails to separate this differentiated pair. $\square$

We are now ready to prove that separating all boundary pairs is sufficient for separating all differentiated pairs. We will prove the contrapositive: if $\mathscr{D}$ fails to separate some differentiated pair $\{P,P'\}$ then it must also fail to separate some boundary pair $\{\hat{P},\hat{P}'\}$. Since $\{P,P'\}$ is differentiated, we have that $t(P) \neq t(P')$. But if $\mathscr{D}$ fails to separate them then $r(P) = r(P')$.

Since every experiment $\mathscr{D}$ produces a convex partition $R_{\mathscr{D}}$ by Proposition 1, there is a path from P to P' entirely in $r(P)$. Since $t(P) \neq t(P')$, there is some first pair of neighbors on this path $\hat{P}$ and $\hat{P}'$ where $t(\hat{P}) \neq t(\hat{P}')$. But since this path lives entirely inside $r(P)$, so $r(\hat{P}) = r(\hat{P}')$. Thus, we have a boundary pair that is not separated, completing the proof. $\square$

Next, we provide two important corollaries. First, recall that testing a restricted model $M = (t_1,\dots,t_n,M_0)$ (where $M_0 \neq \emptyset$) is equivalent to classifying model $M' = (t'_1,t'_2)$ where $t'_1 = \bigcup_i t_i$ and $t'_2 = M_0$. This gives the following corollary.

Corollary 1 (*Characterization of Experiments that Test M*)**.** Experiment $\mathscr{D}$ tests a model $M = (t_1,\dots,t_n,M_0)$ if and only if it separates every pair of neighbors P,P' such that $P \in \bigcup_i t_i$ and $P' \in M_0$.

Finally, an experiment can simultaneously classify and test a restricted model $M = (t_1,\dots,t_n,M_0)$ because doing so is equivalent to classifying the complete model $M' = (t_1,\dots,t_n,t'_{n+1})$ where $t'_{n+1} = M_0$. For this corollary, recall that if $P \in M$ and $P' \in M_0$ then this pair is differentiated by M.

Corollary 2 (*Characterization of Experiments that Test and Classify M*)**.** Experiment $\mathscr{D}$ tests and classifies a model $M = (t_1, \ldots, t_n, M_0)$ if and only if $\mathscr{D}$ separates every pair of neighbors on the permutohedron that are differentiated by M.

V. Classifying Restricted Models

We now focus on classifying a restricted model, which means the researcher wants to identify the subject's type while assuming orders in M_0 cannot be observed. Theorem 1 may not apply in this situation, since it's now possible that a type t_i shares no boundaries with another type t_j in the model. For example, consider $X = \{a,b,c,d\}$ and a model with only two types: those orders for which a is top-ranked, and those for which a is bottom-ranked. Those two types share no neighbors in the permutohedron, and so this model has no boundary pairs.

We can, however, obtain an analogous theorem by working on a restricted permutohedron obtained by removing all rankings in M_0 from the permutohedron. We also remove all edges that contain a ranking in M_0. In doing so, it's possible we completely remove the shortest paths between two rankings P and P'. As we show in Proposition 2 is Section VII, two rankings are separated by an experiment if and only if the experiment contains a set listed on the edges of the shortest path between them. Thus, if *every* shortest path between two rankings is removed from the permutohedron, the information relevant to differentiating the rankings is "lost." To correct this, we reconnect those rankings for which every shortest path between them was deleted. We now formally present this augmented version of the labeled permutohedron.

The set of *restricted neighbors* for M is defined as every pair $P, P' \in M$ such that there does not exist a different $P'' \in M$ along any shortest path between P and P'. The *restricted labeled permutohedron* is a tuple $(\mathscr{P} \setminus M_0, E, L)$, which consists of a graph with nodes $\mathscr{P} \setminus M_0$ and edges E between the set of *restricted neighbors*, along with the edge labels $\tilde{L}$ defined as follows: $\tilde{L}(E) = \left\{S \subseteq X : \mathrm{dom}_P(S) \neq \mathrm{dom}_{P'}(S)\right\}$. That is, the edges are labeled with all the sets for which the neighboring rankings choose differently.

For instance, consider a model where $M_0 = \{adcb, dacb\}$. Its restricted permutohedron is shown in Figure VII. Rankings $adbc$ and $acdb$ are not neighbors in the original permutohedron since they differ by more than one transposition. But there is a unique shortest path between these rankings: $(adbc, adcb, acdb)$. Since $adcb \in M_0$ then $adbc$ and $acdb$ become restricted neighbors. Similarly, $dabc$ and $dcab$ become restricted neighbors, since the only ranking on a shortest path between them is $dacb$, which is in M_0.

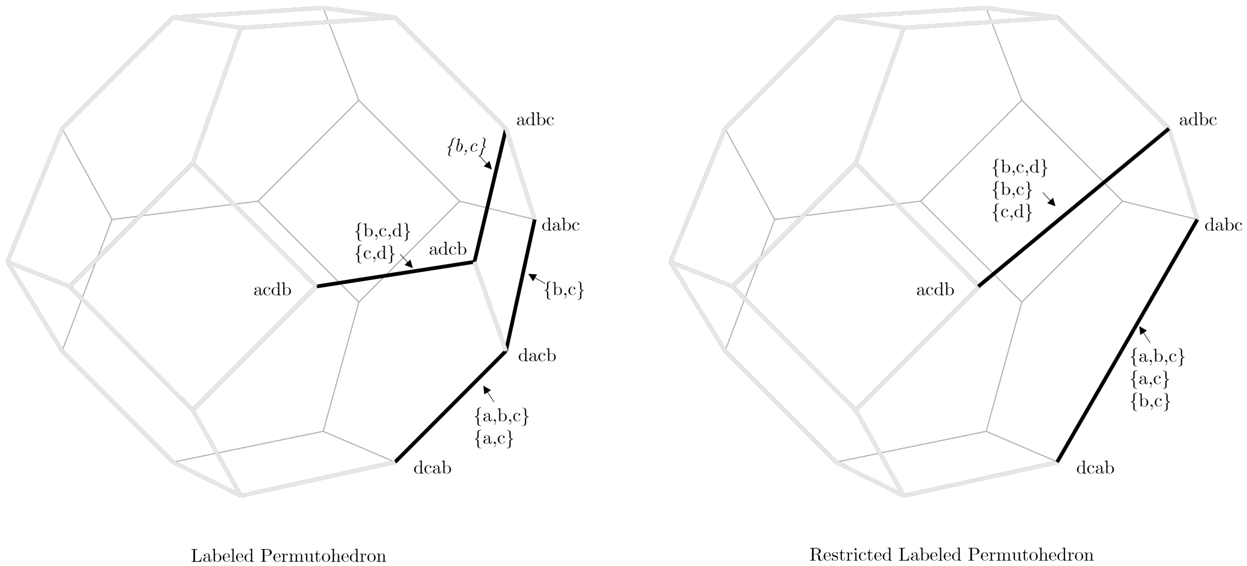

FIGURE VII. The restricted labeled permutohedron for 4 objects $X = \{a,b,c,d\}$ with $M_0 = \{adcb, dacb\}$. (Only the bold edges have been labeled.)

As we will prove below, an analogous result to our Theorem 1 applies to the restricted labeled permutohedron when it comes to classifying restricted models. Perhaps unsurprisingly, the proof of this result is remarkably similar to that of Theorem 1. One complication is that the partition induced by an experiment on the restricted permutohedron is not necessarily convex, a property leveraged in the previous proof.

For instance, suppose we want to classify a restricted model with two types and objects $\{a,b,c,d\}$. The two types are all the rankings with a ranked first and the single ranking $bcda$. In constructing the restricted permutohedron, all shortest paths between each of the rankings with a first (which we denote $a***$) and the ranking $bcda$ are removed. Thus, each of the $a***$ rankings becomes a restricted neighbor of $bcda$.

Now consider the shortest paths between a pair of rankings on opposing corners of the hexagonal face of the unrestricted permutohedron with all the $a***$ vertices: for instance $abcd$ and $adcb$. Any path between this pair that remains on the hexagonal face involves three edges. However, the shortest path on the restricted permutohedron is a two-edge path passing through the vertex $bcda$. Since $bcda$ is not in the same set of the experiment, the experiment is not convex with respect to the shortest paths on the *restricted* permutohedron. This example is depicted in Figure VIII.

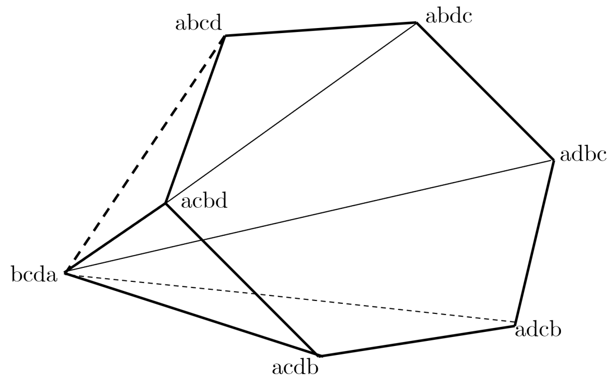

FIGURE VIII. The restricted permutohedron for objects $X = \{a,b,c,d\}$ with $t_1 = \{a***\}$ and $t_2 = \{bcda\}$. Dotted lines show the shortest path between $abcd$ and $adcb$, which passes outside the experiment set containing these two rankings.

However, in the proof of Theorem 1 convexity of the experiment partition was only used to ensure the existence of a path between any two rankings in the same set of the experiment partition that remains in that set. More formally, that the experiment partition is a set of connected subgraphs. We prove this weaker condition within the proof of Theorem 2, though we note that the convexity of the experiment partition on the full permutohedron still plays a key role in this proof.

We are now ready to state and prove a definition and theorem analogous to Theorem 1 for classification of restricted models.

Definition 8 (*Restricted Boundary Pairs*)**.** Fix a model M. A pair $\{P,P'\}$ with $P,P' \in M$ is a *restricted boundary pair* for model M if it is a differentiated pair such that P and P' are restricted neighbors for M.

Theorem 2 (*Characterization of Experiments that Classify Restricted M*)**.** Experiment $\mathscr{D}$ classifies a model $M = (t_1,\ldots,t_n,M_0)$ if and only if $\mathscr{D}$ separates every restricted boundary pair for model M.

Proof of Theorem 2. Necessity is simple: If $\mathscr{D}$ classifies M then *all* differentiated pairs are separated by $\mathscr{D}$, and so every boundary pair must also be differentiated.

For sufficiency, recall that $R_{\mathscr{D}} = (r_1,\ldots,r_k)$ is the partition of $\mathscr{P}$ generated by experiment $\mathscr{D}$. For any model M, define $\tilde{R}_{\mathscr{D}} = (\tilde{r}_1,\ldots,\tilde{r}_k)$ to be the partition of $\mathscr{P} \setminus M_0$ defined by $\tilde{r}_i = r_i \cap (\mathscr{P} \setminus M_0)$ for each i. Before proceeding, we first prove that the sets in $\tilde{R}_{\mathscr{D}}$ are connected subgraphs.

Lemma 5 ($R_{\mathscr{D}}$ *is a Set of Connected Subgraphs*)**.** Each set $\tilde{r}_i$ in $\tilde{R}_{\mathscr{D}}$ is a connected subgraph on the restricted permutohedron.

Proof of Lemma 5. Choose any two rankings P and P' such that $r = r(P) = r(P')$. The proof is by induction on the graph distance between P and P'. If P and P' of distance 1, then they are restricted neighbors and thus connected within the set r. Now suppose they are graph distance d apart, either they are restricted neighbors or there is some vertex on a shortest path between them in the unrestricted permutohedron. Since experiments are convex by Proposition 1, that vertex is in r. Furthermore, that vertex is no more than distance $d-1$ from both P and P'. If every pair of rankings in the same set of the experiment partition that are no more than distance $d-1$ apart are connected within their experiment set, then two rankings in the same set that are distance d are connected as well. □

We are now ready to prove that separating all restricted boundary pairs is sufficient for separating all differentiated pairs. We will prove the contrapositive: if $\mathscr{D}$ fails to separate some differentiated pair $\{P,P'\}$ then it must also fail to separate some boundary pair $\{\hat{P},\hat{P}'\}$. Since $\{P,P'\}$ is differentiated, we have that $t(P) \neq t(P')$. But if $\mathscr{D}$ fails to separate them then $r(P) = r(P')$.

By Lemma 5, there is a path from P to P' entirely in $r(P)$. Since $t(P) \neq t(P')$, there is some first pair of neighbors on this path $\hat{P}$ and $\hat{P}'$ where $t(\hat{P}) \neq t(\hat{P}')$. But since this path lives entirely inside $r(P)$, so $r(\hat{P}) = r(\hat{P}')$. Thus, we have a boundary pair that is not separated, completing the proof. □

VI. Set-Valued Choices

Thus far, we have focused on experiments in which only one object can be chosen from each menu, which we refer to as *choose-one menus*. Experiments that use choose-one menus are both simple and easy to incentivize. A generalization of this allows menus in which subjects choose their top k items. We refer to these as *choose-k menus*. In this case, the subject is paid a lottery in which each of the chosen items is given to the subject with equal probability. This is incentive compatible under the same assumptions as choose-one menus, so long as subjects perceive the lottery probabilities as objective and truly identical (Azrieli et al., 2020).

By including choose-k menus it is possible to reduce the number of decisions needed in a minimal experiment for some models. For instance, consider objects $X = \{a,b,c\}$ and the complete model in which every ordering is in a separate type. This model can be classified using three choose-one menus: $D_1 = \{a,b\}, D_2 = \{a,c\}, D_3 = \{b,c\}$. However, it can be classified with two sets if choose-2 menus are permitted. Asking the subject their favorite choice from $\{a,b,c\}$ and their top two choices from $\{a,b,c\}$ is sufficient to identify the subject's entire rank ordering.

As another example, suppose we permit these choose-k menus in our introductory example with the bundles $a = (2,2), b = (3,1), c = (1,3)$. The model of convexity on these bundles is $t_1 = \{abc, acb, bac, cab\}, M_0 = \{bca, cba\}$. Recall that with choose-one menus, the minimal experiment is $D_1 = \{a, b\}, D_2 = \{a, c\}$. However, if we allow choose-k menus, having subjects choose two objects from $\{a, b, c\}$ is minimal, since this choice can identify that a is not ranked last.

Our results for choose-one menus presented above extend rather naturally to experiments that include choose-k menus. To achieve this, we can expand the edge labels on the permutohedron to include this richer class of sets. In this case, we need to designate not only the set of objects in the menu but also the number of objects to be chosen from that menu.

We adopt the notation of including the number of objects to be chosen after the set of objects and separated by a colon. So, the label $\{a, b, c\} : 2$ indicates that two objects are to be chosen from the set $\{a, b, c\}$. As before, we label each edge with the menus for which the neighboring rankings choose differently. The labeled permutohedron for objects $\{a, b, c\}$ with choose-2 menus included is shown in Figure IX.

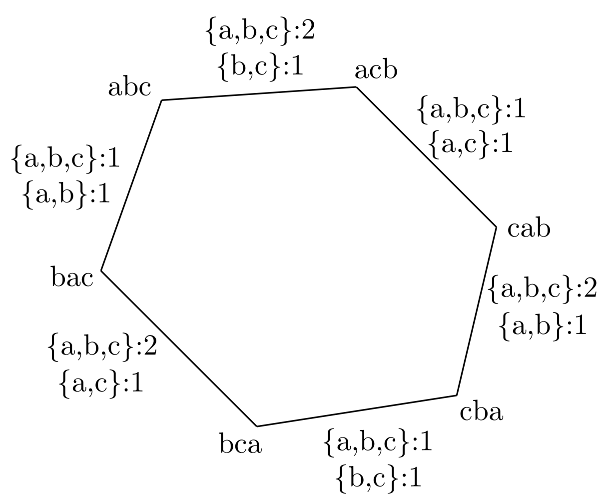

FIGURE IX. The labeled permutohedron for objects $X = \{a, b, c\}$ with choose-2 menus included.

In Appendix IX, we show that our Theorems 1 and 2 can be generalized to include choose-k menus. The proof hinges on the fact that experiments remain convex on this expanded permutohedron- a result leveraged in both of our theorem proofs. Recall that a set is convex on the permutohedron if that set contains all of its shortest paths. In Proposition 1 we prove that the partition created by any experiment using choose-one menus is a convex partition. This proof relies primarily on Lemma 3, which shows that every shortest path between two rankings contains a single instance of each of the pairs of objects for which those rankings choose differently. This is the transposition set $T(P, P')$.

For intuition for why convexity extends to this larger class of experiments, suppose that an experiment including choose-k menus created a non-convex partition. This implies there are two rankings P, P' who make the same choices in the experiment, but for which there is some ranking P'' on a shortest path between P, P' that chooses differently in the experiment. Thus, there must be some menu for which that ranking P'' chooses differently. Since P'' chooses differently, there must be some pair of non-identical objects x and x' such that P and P' include x but not x' in their choice set from the relevant menu, but P'' includes x' but not x. This implies for P and P', $x \succ x'$ but for P'' $x' \succ x$. However, this would imply that the pair $\{x, x'\}$ appears *at least twice* on a shortest path between P and P', violating Lemma 3.

Since, for each edge, including choose-k menus results in edge labels that are a superset of the edge labels with exclusively choose-one menus, there are more options for covering the edges between boundary pairs. This can reduce the number of menus in a minimal experiment.[17]

VII. Properties of Shortest Paths

Recall that a convex set on a graph contains all of its shortest paths. In Proposition 1, we prove that every set in an experiment partition is convex (on the full permutohedron). This plays a key role in our proofs of Theorems 1 and 2. However, given the structure of our proofs, it is easy to overlook the significance that shortest paths play in separating rankings. In this section, we highlight some facts about shortest paths that might provide additional insight into our results and the use of the permutohedron in studying preferences.

As we show below, the labels on any shortest paths are a characterization of the sets that can differentiate two rankings. Furthermore, while there may be multiple shortest paths between two rankings, the collection of sets on those paths are identical. Thus, to differentiate any two rankings, it is sufficient to pick *any* shortest path between the rankings and ensure there is some set on that shortest path included in the experiment.

Take for example the rankings $P = abdc$ and $P' = cabd$. These differ by three transpositions: $T(P, P') = \{\{a, c\}, \{c, b\}, \{c, d\}\}$. Consistent with Lemmas 2 and 3, both shortest paths between the rankings have length three and the three sets in $T(P, P')$ appear exactly once in the labels along the two paths. This is shown in Figure V. Notice that on the two shortest paths: $(abcd, acbd, cabd, cadb)$ and $(abcd, acbd, acdb, cadb)$, the edge labels are identical and include the sets $\{c, b\}, \{a, c\}, \{b, d\}, \{b, c, d\}, \{a, b, c\}, \{a, b, c, d\}$. The two rankings choose differently from each set. For instance, $abdc$ chooses b from $\{c, b\}$ while $cabd$ chooses c. Furthermore, there is *no other set* for which these two rankings choose differently.

[17] When including choose-k menus, choosing the experiment ordering is not as straight-forward when the goal is to minimize the number of subject choices. For instance, the menu $\{a, b, c\} : 2$ could be considered a single choice of two objects from a set of three, or it could be considered two choices; first a choice of one object from $\{a, b, c\}$ and a second choice of one object from whatever pair remains. In this way, the experiment $D_1 = \{a, b, c\} : 2$ might be considered larger than $D_1 = \{a, b\} : 1, D_2 = \{a, c\} : 1$.

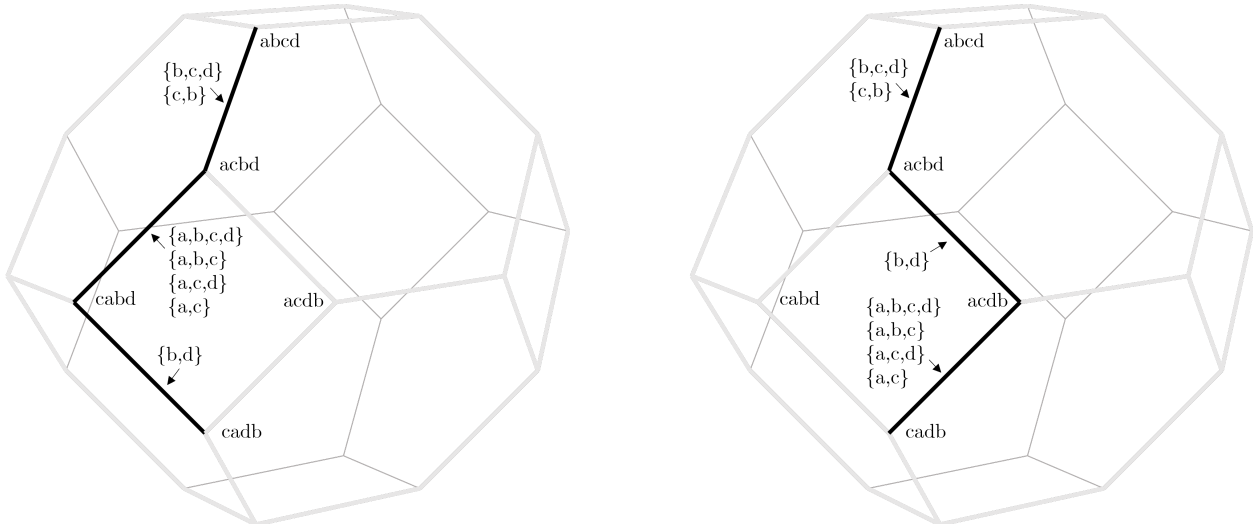

FIGURE X. The Two Shortest Paths from $abcd$ to $cadb$

We now prove these results formally. Most of the groundwork for this result was laid in the lemmas leading to the convexity result in Proposition 1.

Proposition 2 (*Characterization of Separation*)**.** Experiment $\mathscr{D}$ separates P from P' if and only if on some shortest path W between P and P' there is at least one set $D_i \in \mathscr{D}$ such that $D_i \in L(W)$.

Proof of Proposition 2. Suppose $\mathscr{D}$ separates P from P'—meaning there is some $D_i \in \mathscr{D}$ such that $\text{dom}_P(D_i) \neq \text{dom}_{P'}(D_i)$—but no $D_j \in \mathscr{D}$ (including D_i) appears in $L(W)$ for any shortest path between P and P'. Let $W = (P_1, ..., P_n)$ be any shortest path. Then for every P_i along path W with $i < n$, $\text{dom}_{P_i}(D_i) = \text{dom}_{P_{i+1}}(D_i)$ and thus, $\text{dom}_P(D_i) = \text{dom}_{P'}(D_i)$ contradicting $\text{dom}_P(D_i) \neq \text{dom}_{P'}(D_i)$.

Conversely, suppose there is a shortest path W with $D_i \in L(W) \cap \mathscr{D}$ but for every $D_j \in \mathscr{D}$ we have $\text{dom}_P(D_j) = \text{dom}_{P'}(D_j)$. Thus, $\text{dom}_P(D_i) = \text{dom}_{P'}(D_i)$. Since D_i appears along W, there must be a pair of rankings P_i and P_{i+1} such that $\text{dom}_P(D_i) = \text{dom}_{P'}(D_i) = \text{dom}_{P_i}(D_i) \neq \text{dom}_{P_{i+1}}(D_i)$. Let $x = \text{dom}_{P_i}(D_i)$ and $x' = \text{dom}_{P_{i+1}}(D_i)$. The set $\{x, x'\} \in T(P_i, P_{i+1})$ but since $\text{dom}_P(D_i) = \text{dom}_{P'}(D_i)$ $\{x, x'\} \notin T(P, P')$. This contradicts Lemma 3. □

Proposition 3 (*All Shortest Path have Identical Labels*)**.** $L(W) = L(W')$ for every shortest path between P and P'.

Proof of Proposition 3. Suppose otherwise, there is a set $D \in L(W)$ such that $D \notin L(W')$. Let $W' = (P_1, ..., P_n)$. For all $i < n$, $dom_{P_i}(D) = dom_{P_{i+1}}(D)$. Thus, $dom_P(D) = dom_{P'}(D)$. For the rest of the proof, let $x = dom_P(D) = dom_{P'}(D)$. Along W', for every x' such that $x \neq x' \in D$, $dom_{P_i}(\{x, x'\}) = dom_{P_{i+1}}(\{x, x'\})$ and so $dom_P(\{x, x'\}) = dom_{P'}(\{x, x'\})$. Thus, $\{x, x'\} \notin T(P, P')$. By Lemma 3, any set of two objects not in the transposition set of P and P' cannot appear on a shortest path between the pair. Thus, for every shortest path W between P and P' and every x' such that $x \neq x' \in D$ we have $\{x, x'\} \notin L(W)$. However, since $D \in L(W)$, there is some ranking $\tilde{P}$ on W such that $x' = dom_{\tilde{P}}(D) \neq x$. The pair $\{x, x'\}$ must be inverted at least once on W and thus, $\{x, x'\} \in L(W)$- a contradiction. □

VIII. Finding Minimal Experiments Via Linear Programming

Our Theorems 1 and 2 do not provide a minimal experiment directly. Instead, they greatly reduce the complexity of finding minimal experiments by focusing the selection of sets to the edges between boundary pairs. In some cases, finding the minimal experiment after applying the theorems is straightforward. In the case of the convexity example discussed in Section II and shown in Figure II, only one set appears on each of the edges between boundary pairs. However, generally, the application of our theorem leaves several possibilities for experiments that test or classify a model. In those cases an algorithm can be constructed to find the minimal experiment.

Following Section II, we focus here on the goal of minimizing the number and size of the menus used in an experiment. Formally, the *lexicographic size ordering* $>$ is defined by $\mathscr{D} > \mathscr{D}'$ if (1) $\mathscr{D}$ contains more menus than $\mathscr{D}'$, denoted $|\mathscr{D}| > |\mathscr{D}'|$, or (2) $|\mathscr{D}| = |\mathscr{D}'|$ and $\sum_{D \in \mathscr{D}} |D| > \sum_{D \in \mathscr{D}'} |D|$. We will show that identifying the minimal experiment under this particular ordering can be solved as a straightforward integer binary linear program. Extending this result to other similar orderings is straightforward.

The algorithm is broken down into two parts. First, we apply the relevant boundary pair theorems to determine the boundary pairs and the sets on the edges between those boundary pairs. This part depends on whether a restricted model is being classified (applying Theorem 1 or 2). Once the boundary pairs and sets on each edge have been enumerated, the algorithm proceeds to solve the resulting set cover problem by converting it into a linear program. This part is identical whether the model is being tested or not.

Part 1A (Complete Models): Apply the Boundary Pair Theorem

(1) Determine the number of objects in the model, n.
(2) Construct the possible rankings of these objects by finding all permutations of length n.
(3) For each ranking, determine its set in model M.
(4) For each pair of rankings P and P' in different sets in M, count the transpositions $|T(P, P')|$. If the $|T(P, P')| = 1$, rankings are a boundary pair.
(5) For each boundary pair, determine the sets for which the relevant rankings choose differently to construct a list of sets for each boundary pair.

Part 1B (Restricted Models): Apply the Boundary Pair Theorem

(1) Determine the number of objects in the model, n.
(2) Construct the possible rankings of these objects by finding all permutations of length n.
(3) For each ranking, determine its set in model M
(4) For each pair of rankings P, P' that are not in M_0 but are in different sets in M, determine the transpositions: $T(P,P')$. If no other $P'' \notin M_0$ is such that $T(P,P'') \subset T(P,P')$ then P and P' are a boundary pair.
(5) For each boundary pair, determine the sets for which the relevant rankings choose differently to construct a list of sets for each boundary pair.

From here the algorithm can proceed identically for both goals, let $E = (e_1, ..., e_m)$ be the set of boundary pairs and $S = \{S_1, ..., S_l\}$ be the sets appearing on the edges of those boundary pairs. There are m boundary pairs and l total unique sets appearing on those edges. A minimal experiment can be found by choosing from the l sets to minimize an objective under the constraint that at least one set is chosen from each boundary pair. This is a set cover problem, which can be solved by an integer binary linear program. Below, $\mathbb{1}_m$ represents a vector of ones of length m.

Part 2: Set Cover by Linear Programming

(1) Construct a $m \times l$ matrix O such that $O_{(i,j)} = 1$ if set S_j appears on boundary pair i, and $O_{(i,j)} = 0$ otherwise.
(2) Construct a lexicographic cost vector c of length l where $c_j = 1 + \frac{\#(S_j)}{n*l}$.[18]
(3) Solve the resulting set cover problem by integer binary linear programming.

$$\begin{aligned} &\text{Minimize} && c^T \cdot x \\ &\text{subject to} && O \cdot x \geq \mathbb{1}_m \quad \& \quad x \in \{0,1\}^l \end{aligned}$$

(4) Each solution x^* defines a minimal experiment, wherein the minimal experiment includes S_j if and only if $x^*_j = 1$.

Example 1. Consider the goal of classifying and testing the model from Section II given by $t_1 = \{abc, acb\}$, $t_2 = \{bac\}$, $t_3 = \{cab\}$, and $M_0 = \{bca, cba\}$. There are four boundary pairs: $\{bac, abc\}$, $\{bac, bca\}$, $\{cab, acb\}$, and $\{cab, cba\}$. Thus, $m = 4$. The sets on the edge between each boundary pair respectively are $\big\{\{a,b,c\},\{a,b\}\big\}$, $\big\{\{a,c\}\big\}$, $\big\{\{a,b,c\},\{a,c\}\big\}$, and $\big\{\{a,b\}\big\}$. There are three unique sets on these edges, given by $S_1 = \{a,b,c\}$, $S_2 = \{a,b\}$, and $S_3 = \{a,c\}$, so $l = 3$. The matrix O and the vector c are therefore

[18]For this vector, the cost of any set is 1 plus a weighted size of the set. Reducing the selected sets by one set will decrease cost by at least 1. The number of total objects (including repetitions) appearing in the chosen sets can never be more than $n*l$ since n is the number of objects and l is the number of sets on the boundary pairs. Thus, the weight $\frac{1}{n*l}$ ensures the costs are lexicographic, prioritizing the number of sets over set size.

$$O = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \text{and} \qquad c = \begin{pmatrix} 1+\frac{3}{12} \\ 1+\frac{2}{12} \\ 1+\frac{2}{12} \end{pmatrix}.$$

The resulting linear program is minimized at $x = (0,1,1)^\intercal$ which corresponds to minimal experiment $\big\{\{a,b\},\{a,c\}\big\}$. To confirm each relevant edge is covered, note that $Ox = (1,1,1,1)^\intercal$.

IX. Discussion

There are obvious similarities between our approach and that taken by the revealed preference literature. Both are interested in understanding when a model can be tested and when it can't. The difference is that the revealed preference literature typically fixes a certain type of choice menu (for example, linear budget sets) and asks which choices from those menus would be consistent with a given model. This allows for "false positives," where the model is false but is not rejected because the data are not sufficiently rich. Our approach instead searches for choice menus from X such that the resulting data will always be rich enough to avoid false positives. Both approaches avoid "false negatives:" correct models are never wrongly falsified.

To illustrate the difference, consider the following revealed preference theorem, due to Fishburn (1975): Given is k binary menus of the form $D_i = \{p_i, q_i\}$, where each p_i and q_i are simple lotteries, Suppose (without loss) that p_i is chosen in each menu.[19] This vector of choices is consistent with expected utility maximization if and only if there is no probability distribution $\lambda \in \Delta(\{1,\dots,k\})$ over decision problems such that $\sum_{i=1}^{k} \lambda_i p_i = \sum_{i=1}^{k} \lambda_i q_i$. In other words, there is no "first stage" lottery λ such that the compound lottery of λ over $(p_i)_{i=1}^{k}$ and the compound lottery of λ over $(q_i)_{i=1}^{k}$ reduce to the same simple lottery.

In Fishburn's theorem, the choice menus are required to be binary menus, but if the number of menus is small then the experiment may fail to detect violations of expected utility. Our approach instead takes a set of possible lotteries X as fixed and asks which choice menus from X could be used so that, regardless of what data is observed, the researcher will be able to conclude definitively whether expected utility is satisfied on X.

For example, suppose a, b, c, and d are all lotteries, that a, b, and c form the vertices of a triangle in the simplex, and that d is in the interior of that triangle. Expected utility preferences have linear indifference curves and thus would require that d (the interior point) is never ranked first or last; beyond that, all other orderings are permissible. To see how Fishburn's theorem applies, consider the experiment $D_1 = \{a,d\}$, $D_2 = \{a,b\}$, $D_3 = \{b,c\}$.

[19] Fishburn's theorem requires that at least one choice represents a strict preference. In this paper, we assume all preferences are strict.

We take this experiment as fixed; it is not chosen to be optimal in any way. A subject with preference ordering $dabc$ (which violates expected utility since d is ranked first) will choose (d,a,b) from these three menus. The three unchosen items are (a,b,c). Since we can find a vector λ such that $\lambda\cdot(d,a,b)=\lambda\cdot(a,b,c)$, we verify that expected utility is rejected.[20] But a subject with preference $abcd$ (which also violates expected utility) would choose (a,a,b), and there is no λ such that $\lambda\cdot(a,a,b)=\lambda\cdot(d,b,c)$. Thus, this experiment does not identify all expected utility violations over these four options.

Our approach instead demands error-free testing and searches for an experiment rich enough to identify all possible failures. Using our algorithm, we find that the minimal experiment for testing expected utility on these four objects is given by $D_1=\{a,d\}$, $D_2=\{b,d\}$, and $D_3=\{c,d\}$. Any subject who violates expected utility on this domain will either pick d in all three menus or in none of them. And any subject consistent with expected utility would pick d in one or two menus. Thus, this experiment perfectly separates those who violate the model from those consistent with it.

In addition, our method can also be used to classify subjects within a given model. For example, it can be used to find in which range a subject's risk aversion parameter lies. The revealed preference literature typically does not focus on these "type identification" exercises; in most applications type identification is econometric rather than deterministic.

One limitation of our method is it takes as given the set of alternatives X.[21] Definitively testing a model such as expected utility is easy when X contains few elements, but if X is large then minimal experiments may become complex and hard to compute. In that case, it may be worthwhile to choose both which $X'\subseteq X$ to use as the space of alternatives, and which experiment is minimal for X'. When studying expected utility, for example, the space of all lotteries is uncountable. In that case, what finite set of lotteries X' would be sufficient for the experimenter's purpose? Here, false positives (failures to reject the model) become problematic, as compliance with the theory on X' doesn't imply compliance on all of X. Similarly, types on X' are necessarily coarser than those on X, so classification becomes less precise as X' becomes small relative to X. Thus, the size of X' represents a trade-off between the size of the resulting minimal experiment, the fineness of types one can separate, and the frequency of false positives we might expect in the full domain. How to choose X' optimally given these trade-offs remains an interesting and important open question, and one that likely depends on the experimenter's particular objective.

[20]Specifically, if $d=\alpha_1a+\alpha_2b+\alpha_3c$ then $\lambda_1=1/(\alpha_1+2\alpha_2+3\alpha_3)$, $\lambda_2=(\alpha_2+\alpha_3)/(\alpha_1+2\alpha_2+3\alpha_3)$, $\lambda_3=\alpha_3/(\alpha_1+2\alpha_2+3\alpha_3)$.

[21]The revealed preference approach has a similar domain restriction: It takes the experiment $\mathscr{D}$ as given. Consequently, the model can only be tested on the domain $X=\bigcup_{S_j\in\mathscr{D}}S_j$.

Throughout the body of the paper, we focused primarily on minimizing the number of choice tasks asked of the subject. However, this is just one possible ordering over experiments. Our main theorems are not specific to this particular ordering. Other orderings may apply in certain settings. For example, a researcher with a tight budget may want to minimize costs. This can be achieved using our methods by assigning an expected (or maximal) cost to every menu. Experiments can then be ordered based on the average (or maximum) of these menu costs. The labeled permutohedron approach can then be used to identify the cheapest experiment that tests or classifies a given model. Another possible ordering would be based on subjects' privacy. If the experimenter can assign a "privacy cost" to each experiment—or to each menu—then it is possible to order the experiments in terms of their expected privacy loss. Our approach can then identify the experiment that tests or classifies a model with the smallest loss in privacy.[22]

References

Astrachan, O., 2003. Bubble sort: an archaeological algorithmic analysis. ACM Sigcse Bulletin 35, 1–5.

Azrieli, Y., Chambers, C.P., Healy, P.J., 2020. Incentives in experiments with objective lotteries. Experimental Economics 23, 1–29. URL: https://doi.org/10.1007/s10683-019-09607-0, doi:10.1007/s10683-019-09607-0.

Azrieli, Y., Chambers, C.P., Healy, P.J., 2021. Constrained Preference Elicitation. Theoretical Economics 16, 507–538.

Berge, C., 1971. Principles of combinatorics. New York 176.

Fishburn, P.C., 1975. Separation theorems and expected utilities. Journal of Economic Theory 11, 16–34.

Guilbaud, G.T., Rosenstiehl, P., 1963. Analyse algébrique d'un scrutin. Mathématiques et Sciences Humaines 4, 9–33.

Hossain, T., Okui, R., 2013. The binarized scoring rule. The Review of Economic Studies 80, 984–1001.

Rousseas, S.W., Hart, A.G., 1951. Experimental verification of a composite indifference map. Journal of Political Economy 59, 288–318.

Savage, L.J., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783–801.

[22]If the privacy ordering is not linear in the menus, then the minimization problem may not represent a linear programming problem. This may make the optimization problem computationally difficult if X is large.

ONLINE APPENDIX: NOT INTENDED FOR PUBLICATION

APPENDIX A. PROOFS FOR EXTENDED EXPERIMENTS USING CHOOSE-k MENUS

We begin this section by extending our framework to choose-k menus. An *extended experiment* $\mathscr{D}^e$ is a family of tuples consisting of sets $\mathscr{D} = \{D_1, \ldots, D_n\}$ and number of choices $M = m_1, ..., m_n$ from those sets. Typical elements of $\mathscr{D}^e$ are denoted (D_i, m_i). Each element has the property that $D_i \subseteq X$, $m_i < |D_i|$, and $(D_i, m_i) \neq (D_j, m_j)$ for all i and $j \neq i$. The interpretation is that each D_i is a menu from which the subject must choose their top m_i most-preferred elements. We define the following choice function:

$$\mathrm{dom}_P^m(X') = \{C \subseteq X' : |C| = m \wedge (\forall x \in C, y \in X'/C)\ xPy\}.$$

Since all orders are assumed to be antisymmetric, $\mathrm{dom}_P^m(X')$ will always contain m elements. Our definition of *separated pairs* for extended experiments simply adopts this extended choice function:

Definition 9 (*Separation with Extended Experiments*)**.** Fix an extended experiment $\mathscr{D}^e$. Two orders P and P' are *separated by* $\mathscr{D}^e$ (or, $\{P, P'\}$ is a *separated pair*) if there exists some $(D_i, m_i) \in \mathscr{D}^e$ such that $\mathrm{dom}_P^m(D_i) \neq \mathrm{dom}_{P'}^m(D_i)$.

Our definitions of the experiment partition, as well as testing and classifying models using an extended experiment, follow as expected from this modified definition of separated pairs.

Proposition 4 (*Extended Experiments are Convex*)**.** Every extended experiment partition $R_{\mathscr{D}^e}$ is convex.

Proof. Suppose the proposition was false, then there is some set in $R_{\mathscr{D}^e}$ that is non-convex. Thus, some pair of rankings P and P' are such that $P' \in r(P)$ but there is some shortest path W between them that does not remain inside $r(P)$.

There must be some P'' on W such that $r(P'') \neq r(P)$, thus there is some set $(D_i, m_i) \in \mathscr{D}^e$ for which $C = dom_P^{m_i}(D_i) \neq dom_{P''}^{m_i}(D_i) = C''$. However, since $r(P) = r(P')$, $dom_P^{m_i}(D_i) = dom_{P'}^{m_i}(D_i) = C$. Since $C \neq C''$ there is some $x \in C$ and $x' \in C''$. $x \in dom_P^{m_i}(D_i)$ and $x'' \in dom_{P''}^{m_i}(D_i)$. However, $x \notin dom_{P}(D_i)$ $x'' \notin dom_{P''}(D_i)$. Thus, it must be that for P and P', $x \succsim x''$ and for P'', $x'' \succsim x$. Thus x and x'' must be inverted at least twice on the path W and so the set $\{x, x''\}$ appears at least twice in on some shortest path from P to P', contradicting lemma 3.

□

Extending this proposition immediately extends the proof of Theorem 1 simply by replacing instances of choose-one experiments $\mathscr{D}$ with extended experiments $\mathscr{D}^e$. We have included the formal proof below for completeness.

Theorem 3 (*Extension of Theorem 1 to Extended Experiments*)**.** Extended experiment $\mathscr{D}^e$ classifies a complete model $M = (t_1, \ldots, t_n)$ if and only if $\mathscr{D}^e$ separates every boundary pair for model M.

Proof of Theorem 3. Necessity is simple: If $\mathscr{D}^e$ classifies M then *all* differentiated pairs are separated by $\mathscr{D}^e$, and so every boundary pair must also be differentiated.

For sufficiency, note that for any experiment $\mathscr{D}^e$ we can define the partition $R_{\mathscr{D}} = (r_1, \ldots, r_k)$ of $\mathscr{P}$ such that P and P' are in the same partition element if and only if they are not separated by $\mathscr{D}^e$. Let $r(P)$ be the partition element containing order P.

Lemma 6 ($R_{\mathscr{D}^e}$ *Refines* M)**.** If $\mathscr{D}^e$ classifies M then $R_{\mathscr{D}^e}$ is a refinement of M, meaning every $r_i \in R_{\mathscr{D}^e}$ is a subset of some $t_i \in M$

The proof of this lemma is by contradiction: If $R_{\mathscr{D}^e}$ were not a refinement of M then there would be an r_i that intersects two different types t_i and t_j. But then there would be some differentiated pair $P \in t_i$ and $P' \in t_j$ such that $r(P) = r(P') = r_i$, meaning $\mathscr{D}^e$ fails to separate this differentiated pair.

We are now ready to prove the sufficient direction of the theorem. We will prove the contrapositive: if $\mathscr{D}^e$ fails to separate some differentiated pair $\{P, P'\}$ then it must also fail to separate some boundary pair $\{\hat{P}, \hat{P}'\}$. Since $\{P, P'\}$ is differentiated we have that $t(P) \neq t(P')$. But if $\mathscr{D}^e$ fails to separate them then $r(P) = r(P')$.

Since every experiment $\mathscr{D}^e$ produces a convex partition $R_{\mathscr{D}^e}$ by proposition 4, there is a path from P to P' entirely in $r(P)$. Since $t(P) \neq t(P')$, there is some first pair of neighbors on this path $\hat{P}$ and $\hat{P}'$ where $t(\hat{P}) \neq t(\hat{P}')$. But since this path lives entirely inside $r(P)$, so $r(\hat{P}) = r(\hat{P}')$. Thus, we have a boundary pair that is not separated, completing the proof. □

We now extend Theorem 2. This relies critically on the extension of Lemma 5— that the experiment partition on the restricted permutohedron is a set of connected subgraphs. However, this follows immediately from the extension of convexity proved above in 4. The entire proof is included here for completeness.

Theorem 4 (*Extension of Theorem 2 to Extended Experiments*)**.** Experiment $\mathscr{D}^e$ classifies a model $M = (t_1, \ldots, t_n, M_0)$ if and only if $\mathscr{D}$ separates every restricted boundary pair for model M.

Proof of Theorem 4. Necessity is simple: If $\mathscr{D}^e$ classifies M then *all* differentiated pairs are separated by $\mathscr{D}^e$, and so every boundary pair must also be differentiated.

For sufficiency, recall that $R_{\mathscr{D}^e} = (r_1, \ldots, r_k)$ is the partition of $\mathscr{P}$ generated by experiment $\mathscr{D}^e$. For any model M, define $\tilde{R}_{\mathscr{D}^e} = (\tilde{r}_1, \ldots, \tilde{r}_k)$ to be the partition of $\mathscr{P} \setminus M_0$ defined by $\tilde{r}_i = r_i \cap (\mathscr{P} \setminus M_0)$ for each i. Before proceeding, we first prove that the sets in $\tilde{R}_{\mathscr{D}^e}$ are connected subgraphs.

Lemma 7 (*$R_{\mathscr{D}^e}$ is a Set of Connected Subgraphs*)**.** Each set $\tilde{r}_i$ in $\tilde{R}_{\mathscr{D}^e}$ is a connected subgraph on the restricted permutohedron.

proof. Choose any two rankings P and P' such that $r = r(P) = r(P')$. The proof is by induction on the graph distance between P and P'. If P and P' of distance 1, then they are restricted neighbors and thus connected within the set r. Now suppose they are graph distance d apart, either they are restricted neighbors or there is some vertex on a shortest path between them in the unrestrcited permutohedron. Since extended experiments are convex by Proposition 4, that vertex is in r. Furthermore, that vertex is no more than distance $d-1$ from both P and P'. If every pair of rankings in the same set of the experiment partition that are no more than distance $d-1$ apart are connected within their experiment set, then two rankings in the same set that are distance d are connected as well. □

We are now ready to prove that separating all restricted boundary pairs is sufficient for separating all differentiated pairs. We will prove the contrapositive: if $\mathscr{D}^e$ fails to separate some differentiated pair $\{P, P'\}$ then it must also fail to separate some boundary pair $\{\hat{P}, \hat{P}'\}$. Since $\{P, P'\}$ is differentiated we have that $t(P) \neq t(P')$. But if $\mathscr{D}^e$ fails to separate them then $r(P) = r(P')$.

By lemma 7, there is a path from P to P' entirely in $r(P)$. Since $t(P) \neq t(P')$, there is some first pair of neighbors on this path $\hat{P}$ and $\hat{P}'$ where $t(\hat{P}) \neq t(\hat{P}')$. But since this path lives entirely inside $r(P)$, so $r(\hat{P}) = r(\hat{P}')$. Thus, we have a boundary pair that is not separated, completing the proof. □