WHICH EXPERIMENTS TEST A MODEL?[†]

PAUL J. HEALY[*] & GREG LEO[**]

ABSTRACT. Given a model of preferences, what choice data or experiment is sufficient to definitively test whether a subject is consistent with the model or to classify them into some "type" within the model? We characterize all experiments that test or classify a given model using a novel graph-theoretic construction: the labeled permutohedron. We show how these results can be used to identify the smallest experiments that achieve a particular goal.

Keywords: Experimental Design; Model Testing; Incentive Compatibility

JEL Classification: C90, C91, D01.

[†]The authors thank Yaron Azrieli, Gary Charness, Jennifer Pate and seminar attendees for their valuable comments and feedback. This paper was previously circulated under the title "Minimal Experiments."
[*]Dept. of Economics, The Ohio State University; healy.52@osu.edu.
[**]Dept. of Economics, Loyola Marymount University; greg.leo@lmu.edu.

I. INTRODUCTION

Imagine a researcher who wants to study a model of choice under risk. In some cases, the researcher may want to classify subjects into types based on their risk aversion parameter under the assumption that their preferences are consistent with some model such as constant relative risk aversion (CRRA). To do this, the researcher offers subjects a variety of menus of lotteries and identifies their risk parameters based on the observed choices.

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

Both of these goals can be accomplished by observing subjects' entire preference ordering over the relevant lotteries, for example by observing choices over all possible pairs of lotteries. Once a subject's entire ranking is known, the researcher can either classify their CRRA parameter or say whether or not their preferences are consistent 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; subjects can often be classified—and models can be tested—with far less information.

In this paper, we ask for any given model of (rational, deterministic) preferences: What are the sets of choice data/experiments that classify subjects within that model or test the model?[1] Our main results are characterizations of experiments (or, more generally, choice datasets) that successfully classify subjects 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.

Given this characterization, the researcher can then identify the optimal experiment to run from among those that successfully classify or test the model. For example, they may wish to run an experiment that has the fewest choices or has the lowest expected cost. We provide an algorithm for solving this optimization problem along with examples of such "minimal" experiments in Section VIII.

Our characterization is based on the permutohedron, a graph whose vertices correspond to all possible strict rankings of a set of choice objects. Two rankings are connected by an edge if they differ by only one transposition of adjacent objects in the ranking. The labeled

[1]A model can consist of a collection of several axioms, such as Savage's subjective expected utility theory, or each axiom on its own could be viewed as a separate model. Or it may not be axiomatic at all. We only require that the model assumes subjects' choices are deterministic and consistent with a complete, reflexive, transitive, and antisymmetric ranking over a finite number of alternatives.

permutohedron augments this by labeling each edge with those menus from which the connected rankings choose differently; see Figure I for an example with three choice objects. The use of (unlabeled) permutohedra to visualize the possible rankings of objects dates back to at least Guilbaud and Rosenstiehl (1963) and seems to have been independently discovered by Kemeny (1959) and Schulman (1979), among others. The mathematics of permutations has roots in Hardy et al. (1934). The mathematical structure of the permutohedron was described by Gaiha and Gupta (1977), and Yu et al. (2019) survey various methods of illustrating rankings, measuring the distance between them, testing various features of a collection of rankings, and aggregating multiple rankings.

Our framework assumes that subjects have a well-defined deterministic preference relation over a finite set of alternatives and that choices from a menu are always consistent with that ranking. A model is simply a partition of the set of all possible rankings. Although this can accommodate many behavioral biases such as violations of stochastic dominance, failures of contingent reasoning, or other-regarding preferences, it does not allow for irrational choice patterns such as those that violate transitivity or the weak axiom of revealed preference. Our framework can handle stochastic choice, but only in a very limited way: If the choice objects themselves are lotteries over alternatives and if stochastic choice is consistent with maximizing some preference ordering over lotteries, then our framework can apply. Most stochastic choice models, however, do not fit this paradigm and therefore are beyond the scope of this paper.[2] In the Discussion section, however, we discuss how one could use our framework to identify whether subjects' choice are stochastic, even if it cannot be used to test or classify models of stochastic choice.

Finally, we assume strict preferences for two reasons: First, the literature on incentive compatible experiments with weak preferences is not well developed; see Azrieli et al. (2018) for a discussion. Second, we focus on choice-from-sets experiments, which cannot distinguish between strict and weak preferences, and thus would not be useful in perfectly classifying or testing a model that allows for indifference. Extending our analysis to include indifferences and, therefore, more sophisticated elicitation techniques would be a useful direction for future work.

Related Research

Our focus in this paper is communicating a new framework for analyzing experimental design. We believe this work provides the first steps in establishing experimental design as

[2]Indeed, when choice is stochastic it's not even obvious how to define incentive compatibility of an experiment (meaning, whether it reveals the underlying stochastic choice function truthfully), or what experiments could be used to perfectly identify their stochastic choice function. Furthermore, our techniques would likely not apply since the space of possible preferences is finite—and can therefore be represented via the permutohedron—but the space of possible stochastic choice functions is uncountable, even when the number of objects is finite.

a formal theoretical problem. Specifically, our paper studies the problem of using limited choice data to test and classify models. A similar vein of literature focuses on "completing" rankings that arise from incomplete choice data. This often involves "fuzzy" preferences, which give not only a ranking but also an intensity (Alonso et al., 2008; Chiclana et al., 2009). Similarly, the conjoint analysis literature (Luce and Tukey, 1964; Green and Rao, 1971; Green and Srinivasan, 1978) aims to estimate preferences from surveys when choice objects have well-defined attributes. In contrast, we use the permutohedron to understand how choices from various menus help identify which preferences are consistent with those choices and therefore how to use experiments to classify subjects into types (defined as sets of preferences) based on their choices.

In Section VIII, we also discuss how our results can be leveraged to find *minimal experiments* which use the fewest choices to test or classify a model. There is a large literature in statistics on the optimal design of experiments under various criteria, where an experiment is used to help estimate the parameter of some data generating process. That literature largely assumes the parameters to be real-valued, such as the slope and intercept of a linear regression (Kiefer, 1959; Atwood, 1969; Smucker et al., 2018; Pukelsheim, 2006). In our setting the parameter of interest is a categorization of the subject's true preference ordering, which is not real-valued; to our knowledge this literature has not yet studied such a scenario.

Outline

In Section II, we demonstrate the framework and 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 extend our results further by showing how they apply to experiments where subjects can choose more than one option from a given menu. In Section VII we explore additional properties of the permutohedron that might be useful in future work. In Section VIII we discuss a practical application of our results: finding experiments that test / classify a model using the fewest choices. Section IX concludes with future directions and applications of our approach.

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 consider each assumption to be a separate *model* of preferences. In this section, we begin

by demonstrating how our methods can be used to characterize the experiments that test and classify various models in this context. Although simple, these examples demonstrate many of our key results and the general intuition they provide scale to larger, more complex models.

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? Of course, 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.

How else can we test this model? One way is to offer the subject a menu of all three plates 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. Another option is to use the menus $D_1 = \{a,b\}$, $D_2 = \{a,c\}$. If a subject chooses a in both menus, the model is validated; otherwise, it fails.

In fact, any experiment that tests this model must contain the menu $\{a,b,c\}$ or the pair $\{a,b\}$ and $\{a,c\}$. Otherwise, there is no guarantee that the experiment will reveal whether the subject ranks a first. Thus, we have a characterization, an experiment tests this model if and only if it includes the menu $\{a,b,c\}$ or includes both menus $\{a,c\}$ *and* $\{a,b\}$.[3]

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 of preferences 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

[3]The experiment can include additional menus as well, but they are not necessary.

condition is $M = \{abc, acb, bac, cab\}$ and the complementary set of rankings outside the model is $M_0 = \{bca, cba\}$.

Unlike the monotonicity example, this model cannot be tested using the choice of a favorite plate from the single menu $D_1 = \{a, b, c\}$. For example, preference ordering bac is convex, but bca is not. Yet, two subjects with these preferences make the same choice from the menu $\{a, b, c\}$ and are not separated by this experiment.

However, the experiment consisting of the two menus $D_1 = \{a, b\}$ and $D_2 = \{a, c\}$ is sufficient to test the model. If a subject chooses a in at least one menu, the model is validated; otherwise, it fails. In fact, any experiment that tests this model *must* contain the two menus $\{a, b\}$ and $\{a, c\}$. If one of these menus is absent, there is no guarantee that the experiment will reveal whether the subject ranks a last. An experiment tests this model if and only if it includes the two menus $\{a, b\}$ and $\{a, c\}$.

Model 3: A General Notion of Types

A model M may further partition the preferences into "types." For example, suppose the researcher is also interested in splitting the convex preferences from the last example 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.[4]

Interestingly, it is possible to classify subjects into these types and simultaneously test the model using the same experiment that is sufficient to test the convexity of preferences for these options: $D_1 = \{a, b\}$ and $D_2 = \{a, c\}$. Subjects of type t_1 will pick (a, a) (meaning a from D_1 and a from D_2), subjects of type t_2 will pick (b, a), and subjects of type t_3 will pick (a, c). Subjects with non-convex preferences (in M_0) will pick (b, c). Thus, this experiment both tests the model and classifies subjects into types within the model.

Since the menus $\{a, b\}$ and $\{a, c\}$ are sufficient to test and classify this model, and the inclusion of these two remains necessary to test convex preferences in the first place, the characterization is the same as the previous example. An experiment tests and classifies this model if and only if it contains $\{a, b\}$ and $\{a, c\}$.

Suppose we wanted to assume that the non-convex preferences were impossible and *only* classify subjects among the types $t_1 = \{abc, acb\}$, $t_2 = \{bac\}$, $t_3 = \{cab\}$ (without also testing the model). We refer to this as classifying a *restricted model*. Again, the experiment $D_1 = \{a, b, c\}$ is sufficient. So too is $D_1 = \{a, b\}, D_2 = \{a, c\}$. In fact, an experiment classifies this

[4]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 convex preference 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.

restricted model if and only if it contains the menu $\{a,b,c\}$ or both of the menus $\{a,b\}$ and $\{a,c\}$.

The Permutohedron

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 depend 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.

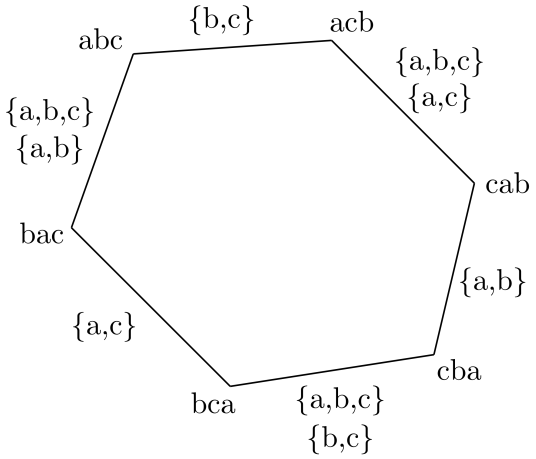

FIGURE I. The labeled permutohedron for three objects.

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 transposition 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 menu $\{b,c\}$, so we label the edge between abc and acb with the menu $\{b,c\}$. 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 which menus separate the non-adjacent rankings. Specifically, our main theorem shows that to test or classify a model, an experiment must contain at least one menu from the edge between every "boundary pair" of rankings: adjacent rankings that belong to different sets in the model.

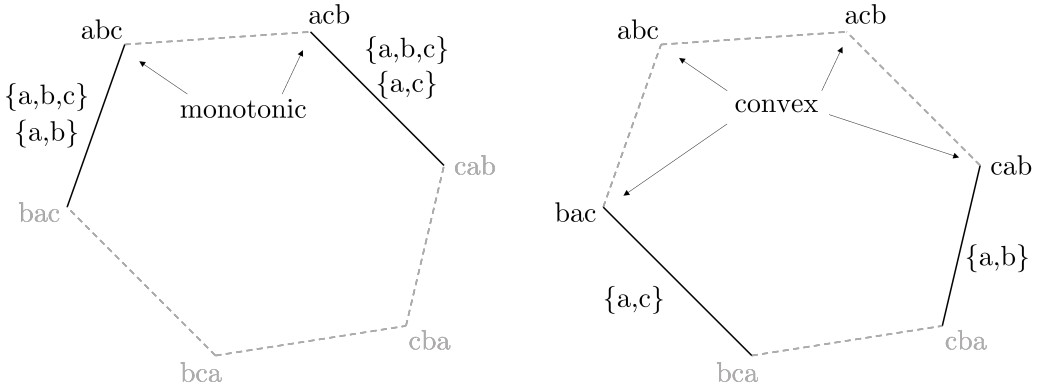

FIGURE II. Boundary pairs for the Model 1, *Monotonic Preferences* (left panel) and Model 2, *Convex Preferences* (right panel). Solid edges are between boundary pairs.

Model 1 (*Monotonic Preferences*). The left panel of Figure II highlights the boundary pairs for the monotonicity example discussed above. The edges between boundary pairs are shown in bold. Here, 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. This leads to the characterization presented above. An experiment that tests this model must contain $\{a, b, c\}$ or both of $\{a, b\}$ and $\{a, c\}$.

Model 2 (*Convex Preferences*). The right panel of Figure II highlights the boundary pairs for the convex preferences example discussed above. There are again two boundary pairs: $\{bac, bca\}$ and $\{cab, cba\}$. Since the first edge contains only the menu $\{a, c\}$, it must be included in the experiment. The second edge contains only the menu $\{a, b\}$, so it also must be included in the experiment. Thus, an experiment tests this model if and only if it includes these two menus.

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 menu D_i. For instance, every experiment that tests the convex preference model discussed above requires subjects to choose from the menus $\{a, b\}$ and $\{a, c\}$. 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 two favorite options (or, equivalently, to eliminate their least favorite option) from the menu $\{a,b,c\}$.

Testing and classifying a model with more types works similarly: the boundary pairs are all of the adjacent pairs that are either in different types (when classifying), or where one is in M and one is in M_0 (when testing).

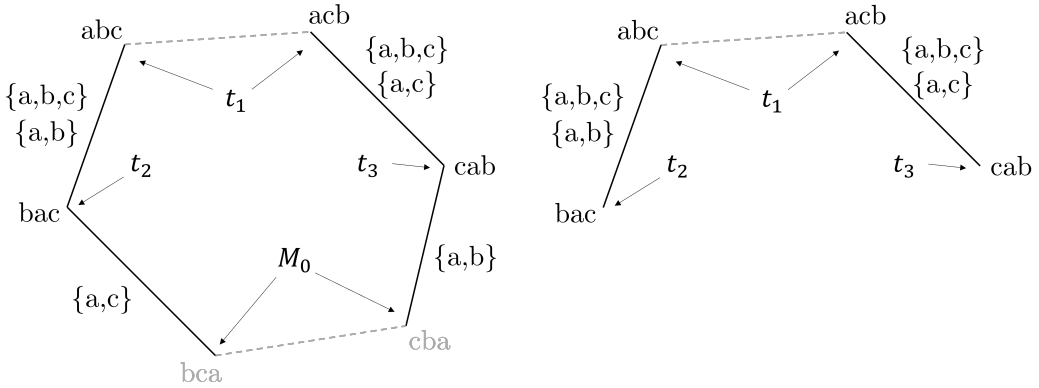

FIGURE III. Boundary pairs for example Model 3. The left panel shows the boundary pairs (solid edges) for testing and classifying the model. The right panel shows the boundary pairs on the restricted permutohedron for classifying only.

Model 3 (*Testing and Classifying*). The left panel of Figure III highlights the boundary pairs for testing and classifying Model 3 discussed above. The edges between boundary pairs are shown in bold. There are four boundary pairs, $\{abc, bac\}$, $\{acb, cab\}$, $\{bac, bca\}$, $\{cab, cba\}$. An experiment will test this model if and only if it contains a menu from each of the four edges connecting these boundary pairs. To cover the edges between $\{bac, bca\}$ and $\{cab, cba\}$, the experiment must contain menus $\{a,b\}$ and $\{a,c\}$. However, these two menus also cover the remaining edges between boundary pairs $\{abc, bac\}$ and $\{acb, cab\}$, so there are no further requirements. This leads to the characterization presented above, an experiment tests and classifies this model if and only if it contains $\{a,b\}$ and $\{a,c\}$.

Additional complications arise when the model being classified (but not tested) does not include all possible preferences. In Section V we demonstrate that an experiment classifies such a "restricted" model (those which do not include every possible ranking) if and only if it contains menus from every boundary pair on a modified graph we call the restricted permutohedron.

Complications for constructing the restricted permutohedron can arise in some models (see Section V for details). Here, however, the restricted permutohedron is constructed simply by deleting the rankings outside the model and their associated edges from the full labeled permutohedron.

Model 3 (*Classifying*). The right panel of Figure III shows the restricted permutohedron for Model 3. The edges between boundary pairs are shown in bold. There are two boundary pairs: $\{abc, bac\}$, $\{acb, cab\}$. Covering these edges can be accomplished either with the inclusion of $\{a, b, c\}$ or the inclusion of both $\{a, b\}$ and $\{a, c\}$, formalizing the characterization presented above.

While these examples are simple, the logic generalizes to any model over a finite set of alternatives X. In the next section, we generalize this theory.

III. The Framework

Let X be a finite set of alternatives, with typical elements denoted by a, b, c, and so on.[5] 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.[6] 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 the model. 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. Overloading notation, we also use M to denote $\cup_{i=1}^{n} t_i$, the set of all orderings in the model. Thus, we can write $P \in M$ for any $P \notin M_0$. 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; we set $t(P) = M_0$ if $P \in M_0$.

An *experiment* is a family of sets $\mathscr{D} = \{D_1, \ldots, D_{|\mathscr{D}|}\}$ such that $D_i \subseteq X$ and $|D_i| \geq 2$. 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 $t(P) \neq t(P')$.

[5] We model X as exogenous, but note in Section IX that the experimenter may first select X from some larger set of possible alternatives. We do not study that selection process in this paper, though it is a very interesting avenue for future research.

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

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 $\text{dom}_P(D_i) \neq \text{dom}_{P'}(D_i)$.[7]

In other words, a model differentiates two orders if the orders belong to different types, while an experiment separates two orders if those orders lead to different choices being observed in at least one menu.

We can now give our main definitions of classifying and testing models using an experiment.

Definition 3 (*Classifying*)**.** An experiment $\mathscr{D}$ *classifies subjects according to model* M (or, more simply, *classifies* M) if every $P \in M$ and $P' \in M$ that are differentiated by M are also 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 the subject's choices in the experiment 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 subject's choices inform the experimenter whether their preference P is included in $M = \cup_{i=1}^{n} t_i$ or belongs to M_0.[8]

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 instead only interested in learning whether $P \in M$, and not in learning the subject's type. An experiment *tests and classifies* a model if it accomplishes both.

To understand this difference, think of the experiment $\mathscr{D}$ as generating a partition of preference orderings based on the possible choices the subject could make in the experiment. Formally, given experiment $\mathscr{D}$, let $R_{\mathscr{D}} = \left(r_1, \ldots, r_q\right)$ be the partition of $\mathscr{P}$ such that P and P' are in the same partition element if and only if they are not separated by $\mathscr{D}$ (meaning $\text{dom}_P(D_i) = \text{dom}_{P'}(D_i)$ for every $D_i \in \mathscr{D}$). Let $r(P)$ be the partition element that contains P. We refer to $R_{\mathscr{D}}$ as the *experiment partition* for experiment $\mathscr{D}$.

For an experiment $\mathscr{D}$ to test a model M, each $r(P)$ must be a subset of either M or M_0. In other words, $R_{\mathscr{D}}$ is a refinement of the two-element partition $\{M, M_0\}$. For an experiment $\mathscr{D}$ to classify M, each $r(P)$ must be a subset of $t_i \cup M_0$ for some $t_i \in M$, or else $r(P) \subseteq M_0$. If $r(P) \subseteq t_i \cup M_0$ then the experimenter (who is assuming M_0 is impossible) classifies the

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

[8] Thus, a model in which $M_0 = \emptyset$ cannot be tested because it incorporates all possible preferences.

subject as type t_i. In this case, they cannot learn whether $P \in t_i$ or $P \in M_0$, meaning they cannot learn whether or not their assumption is true. If $r(P) \subset M_0$ then they do learn that their assumption is false and are therefore unable to classify any subject with such a preference. Combining these insights gives an alternative definition of classifying a model: experiment $\mathscr{D}$ classifies model M if $R_{\mathscr{D}}$ is a refinement of $\{t_1, \ldots, t_n\}$ *after* all orders in M_0 are removed from $R_{\mathscr{D}}$.[9] An experiment $\mathscr{D}$ tests and classifies model M if $R_{\mathscr{D}}$ is a refinement of $(t_1, \ldots, t_n, M_0)$.

As an example of classifying without testing, consider the Holt and Laury (2002) experiment widely used to estimate CRRA risk aversion parameters. Subjects make ten binary choices between a low-risk and a high-risk lottery. The binary menus are presented as a 10-row list in which the difference in expected values between the lotteries is decreasing down the list, meaning a subject with CRRA preferences working down the list will switch only once from choosing the low-risk lottery to the high-risk lottery. If ρ represents the CRRA risk aversion parameter, then a subject who switches immediately has $\rho < -0.95$, a subject who switches after the second row has $\rho \in (-0.95, -0.49)$, and so on. Someone who never switches has $\rho > 1.37$. What model does this classify? It is the model in which t_1 contains all CRRA orderings with $\rho < -0.95$, t_2 contains all CRRA orderings with $\rho \in (-0.95, -0.49)$, and so on, up to t_{10}, which contains all CRRA orderings with $\rho > 1.37$. And M_0 contains all non-CRRA preferences, including those that violate expected utility altogether. But the Holt and Laury (2002) experiment does not test this model; a subject with non-CRRA preferences might still exhibit a single switch point, in which case the experimenter would wrongly classify them as belonging to some set t_i inside the model. Formally, there exists some $P \in M_0$ and type t_i in the model such that $r(P) \cap t_i \neq \emptyset$.

Even though the Holt-Laury experiment does not test the CRRA model, it is still possible for the model to be disproven via this experiment. This happens when a subject exhibits "multiple switching" behavior, switching back to the low-risk lottery after picking a high-risk lottery on an earlier row. These choices reveal that CRRA is violated, meaning $P \in M_0$. Multiple switching occurs in a small but notable fraction of subjects across studies. Such data are obviously problematic for experimentalists who are implicitly assuming that all subjects have CRRA preferences; multiple switching data are often dropped, or else a single switch point is imputed from the range of observed switches, effectively forcing conformity with the model. Many researchers simply sweep the problem under the rug by forcing subjects to switch only once, so that for every P there is some t_i for which $\emptyset \neq (r(P) \setminus M_0) \subseteq t_i$.

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 $M = (t_1, \ldots, t_n, M_0)$

[9] Formally, for each $P \in M$ let $r^M(P) = r(P) \cap M$ and then define $R^M_{\mathscr{D}} = \{r^M(P)\}_{P \in M}$ to be the resulting partition of M. Experiment $\mathscr{D}$ is said to classify model $M = \{t_1, \ldots, t_n, M_0\}$ if $R^M_{\mathscr{D}}$ is a refinement of $\{t_1, \ldots, t_n\}$.

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, the 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.[10]

The Permutohedron

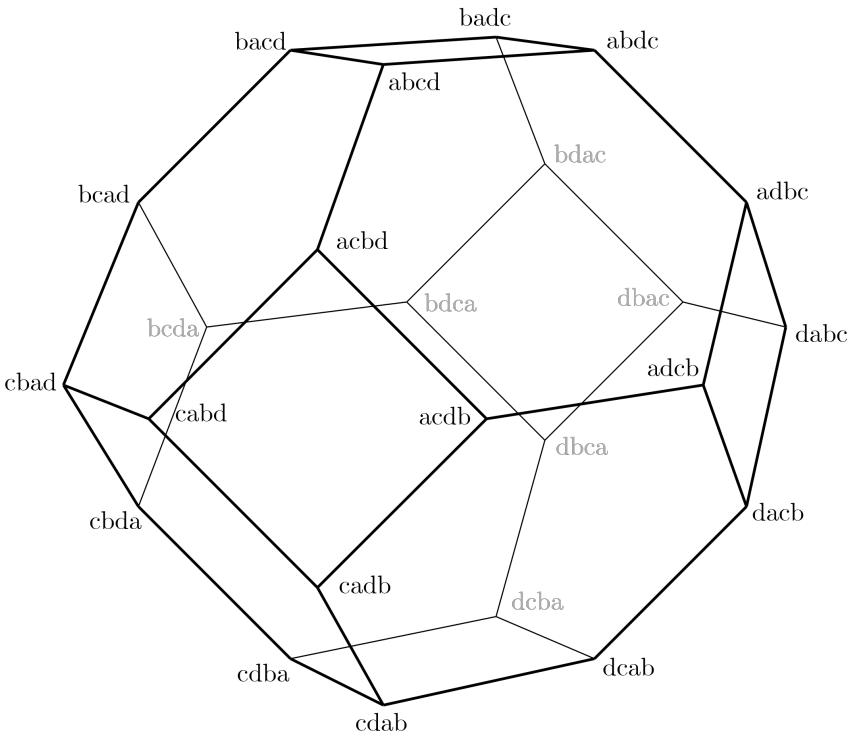

FIGURE IV. The permutohedron for four objects $X = \{a, b, c, d\}$

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:

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

We say P and P' are *neighbors* if $|T(P, P')| = 1$. $|T(P, P')|$ is known as the Kendall tau distance between the rankings P and P' (Kendall, 1938, 1948).[11] The *transposition graph* is a

[10] In this paper, we take a model M as fixed and determine which experiments test/classify that model. The reverse of this problem, determining which models can be tested/classified by a given experiment $\mathscr{D}$, is straightforward. Each experiment induces an experiment partition $R_{\mathscr{D}}$. This is the finest model that experiment $\mathscr{D}$ can test and classify. Because of this, any model M that is coarser than $R_{\mathscr{D}}$ can also be tested and classified by $\mathscr{D}$.

[11] This distance metric was also suggested by Kemeny (1959).

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*. Since the sum of the coordinates is fixed for any ranking, the permutohedron is usually displayed in the $|X|-1$ dimensional simplex; for example, the permutohedron for four objects is shown as a three-dimensional shape in Figure IV, while the three-object permutohedron appears as a hexagon in Figure I.[12]

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 edge $E = \{P,P'\} \in \mathscr{E}$, $L(E) = \{S \subseteq X : \text{dom}_P(S) \neq \text{dom}_{P'}(S)\}$. 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.

There is a simple and useful characterization of the set of labels $L(E)$ on the edge between two neighbors. If P and P' are neighbors, then they differ only in their ranking of two adjacent objects x and x', meaning $T(P,P') = \big\{\{x,x'\}\big\}$. If for any $S \subseteq X$ we define $B(S;P) = S \cup \{x \in X : (\forall y \in S)\ yPx\}$ to be the objects in X that are either in S or are ranked worse than everything in S according to P, then we must have that $B\big(\{x,x'\},P\big) = B\big(\{x,x'\},P'\big)$. Thus, any set D_i will separate P and P' (meaning $D_i \in L(\{P,P'\})$ if and only if (1) D_i contains $\{x,x'\}$, and (2) $D_i \subseteq B\big(\{x,x'\},P\big)$.[13] Thus, $L(\{P,P'\}) = \big\{S \subseteq X : \{x,x'\} \subseteq S \subseteq B(\{x,x'\},P)\big\}$. This also helps enumerate the number of sets in $L(E)$: There are $|B(\{x,x'\},P)| - 2$ objects ranked strictly worse than x and x' according to P (and P'). Thus, the number of menus on the edge between these rankings is $2^{|B(\{x,x'\},P)|-2}$.

A *path* W between P and P' is a finite sequence of v adjacent nodes $(P_1 ..., P_v)$ with $P_i \neq P_j$ for $i \neq j$ such that $P_1 = P$, $P_v = P'$ and $\{P_i,P_{i+1}\} \in \mathscr{E}$. The *length* of path W is defined as $v-1$ (the number of edges). 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' with a smaller length. Shortest paths may not be unique.

[12] 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).

[13] We thank a referee for pointing out this useful fact.

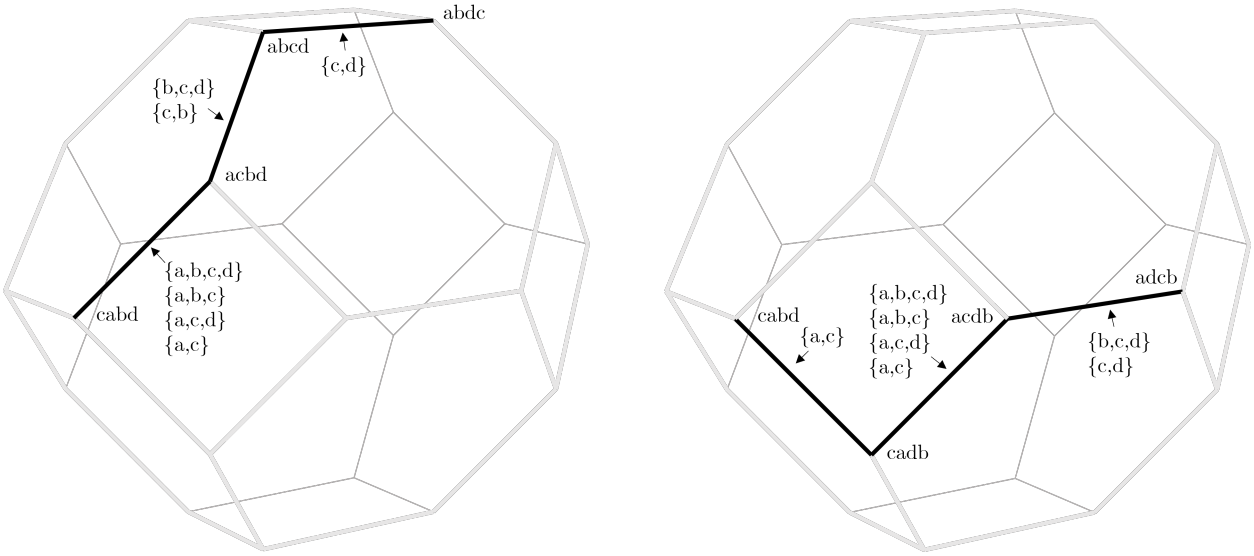

FIGURE V. The shortest path from $abcd$ to $cabd$ and one of the two shortest paths from $cabd$ to $adcb$. The edges have been labeled along each path.

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.

As can be seen in Figure VI, there is a lot of structure in the way 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).[14] This implies that each

[14]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.

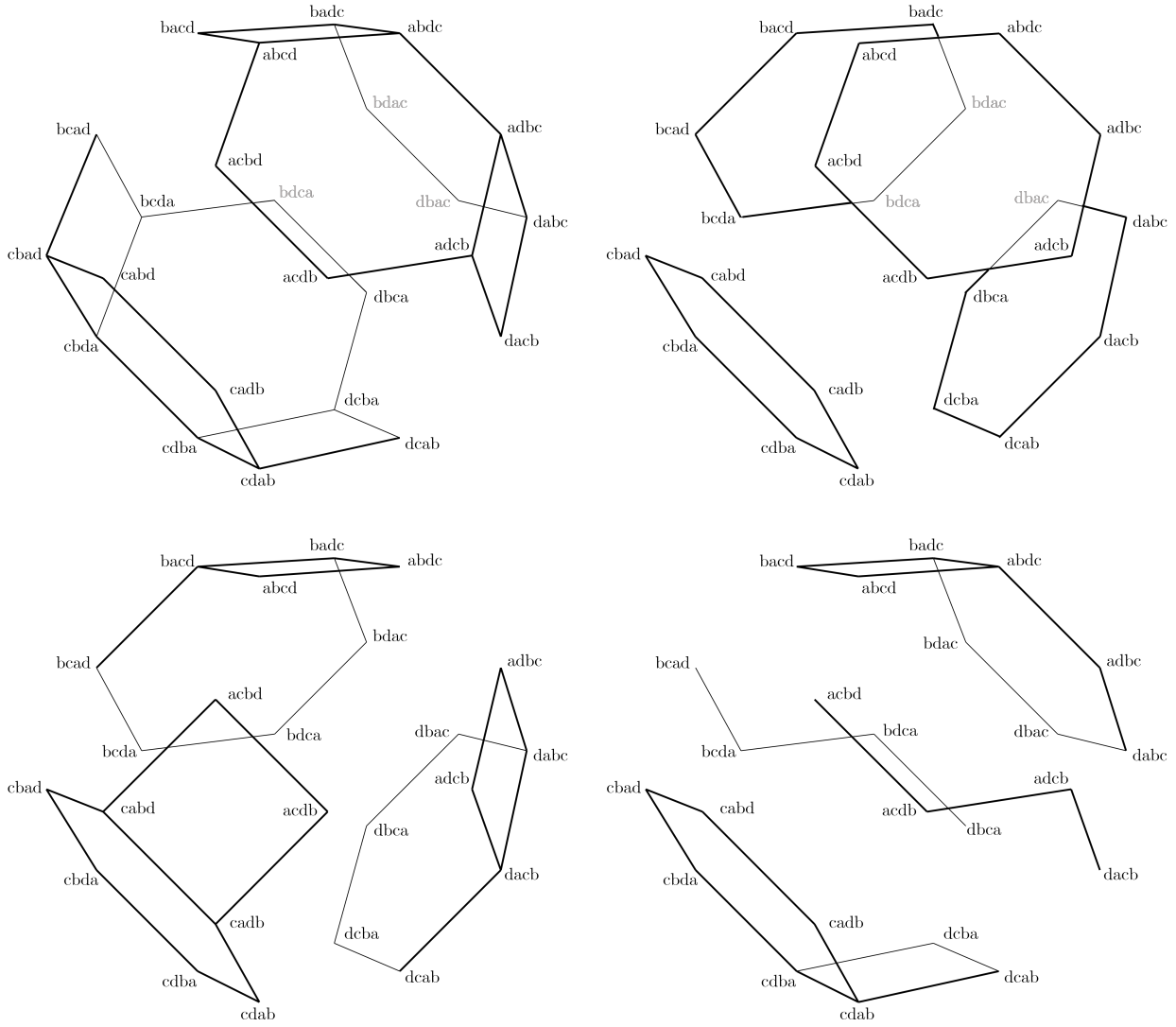

FIGURE VI. Induced graphs for experiments (clockwise) $\mathscr{D} = \{\{a,c\}\}, \mathscr{D} = \{\{a,b,c,d\}\}, \mathscr{D} = \{\{b,c,d\}\}, \mathscr{D} = \{\{a,c\},\{c,b\}\}$.

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 a higher-dimensional permutohedron. The five-object permutohedron, for instance, contains five subgraphs isomorphic to the four-object permutohedron shown in Figure IV.

The fact that experiments generate convex partitions was proven by Azrieli et al. (2021), though the insight is also apparent in Lambert (2019). Azrieli et al. (2021) provide a simple geometric proof by converting each ranking into a convex set of utility vectors in $\mathbb{R}^n$ consistent with that ranking, and then showing that an experiment generates a union of such sets that itself must be convex. We provide here a new proof that is entirely graph-theoretic, given our definition of convexity is in terms of shortest paths.

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. Without loss of generality let $x_1,x_2,...,x_t$ be a sequence of objects that are adjacent in P such that x_iPx_{i+1} with $\{x_1,x_t\} \in T(P,P')$. By assumption, since no adjacent pair in P is transposed in P', and since $\{x_1,x_t\} \in T(P,P')$, we have $x_1P'x_2P'\dots P'x_tP'x_1$, 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 Lemma 1 such a pair will always exist as long as $P_i \neq P'$, and because only adjacent transpositions are made, $T(P_{i+1},P') \subset T(P_i,P')$. Thus, the sequence transforms P into P' with $|T(P,P')|$ adjacent transpositions.[15] □

Since any 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 $S \notin L(W)$.

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

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 has only one set on its label with $|S| = 2$, every set $S \in T(P,P')$ 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. Thus, 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

Our first theorem characterizes when an experiment $\mathscr{D}$ classifies a complete model $M = (t_1,\ldots,t_n)$. To understand the result, recall the partition of preferences induced by the experiment $\mathscr{D}$, which we call the *experiment partition*. It is the partition $R_{\mathscr{D}} = \{r_1,\ldots,r_q\}$ of $\mathscr{P}$ such that P and P' are in the same partition element if and only if P and P' would make the same choices in every menu $D_i \in \mathscr{D}$; in our terminology this means that P and P' are not separated by $\mathscr{D}$ (see Definition 2). Again, $r(P)$ is the experiment partition element containing order P, and recall that $t(P)$ is the type (model partition element) containing P. If $\mathscr{D}$ successfully classifies model $M = (t_1,\ldots,t_n)$ then it must be the case that $r(P) \subseteq t(P)$ for every P. Mathematically, this means that $R_{\mathscr{D}}$ is a refinement of M. Intuitively, it means that the experiment collects at least as much information about P as is necessary to identify to which type it belongs in M. This is summarized by the following lemma, which is key to our main result below.

Lemma 4 ($R_{\mathscr{D}}$ *Refines* M)**.** If $\mathscr{D}$ classifies M then the experiment partition $R_{\mathscr{D}}$ 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. This follows immediately from the definition of classifying a model: 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. □

We can rephrase the idea of $R_{\mathscr{D}}$ being a refinement of M by using the language of differentiated pairs. Recall that $\{P, P'\}$ is a differentiated pair if P and P' are assigned to different types in the model. 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*.

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*)**.** The following are equivalent:

(1) Experiment $\mathscr{D}$ classifies a complete model $M = (t_1, \ldots, t_n)$,
(2) $\mathscr{D}$ separates every boundary pair for model M, and
(3) The experiment partition $R_{\mathscr{D}}$ is a refinement of the model partition M.

Proof of Theorem 1. Equivalence between (1) and (3) follows immediately from definitions, so we focus on proving that (1) if and only if (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 separated.

For sufficiency, we will use Lemma 4 to 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. □

Next, we provide two important corollaries. First, recall that testing a restricted model $M = (t_1, \ldots, t_n, M_0)$ (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*)**.** The following are equivalent:

(1) Experiment $\mathscr{D}$ tests a model $M = (t_1, \ldots, t_n, M_0)$,
(2) $\mathscr{D}$ separates every pair of neighbors P, P' such that $P \in \bigcup_i t_i$ and $P' \in M_0$, and
(3) The experiment partition $R_{\mathscr{D}}$ is a refinement of the two-element partition $(\bigcup_i t_i, M_0)$.

Finally, an experiment can simultaneously classify and test a restricted model $M = (t_1, \ldots, t_n, M_0)$ because doing so is equivalent to classifying the complete model $M' = (t_1, \ldots, t_n, t'_{n+1})$ 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*)**.** The following are equivalent:

(1) Experiment $\mathscr{D}$ tests and classifies a model $M = (t_1, \ldots, t_n, M_0)$,
(2) $\mathscr{D}$ separates every pair of neighbors on the permutohedron that are differentiated by M, and
(3) The experiment partition $R_{\mathscr{D}}$ is a refinement of the $n+1$-element partition $(t_1, \ldots, t_n, M_0)$.

We motivated these theorems by imagining a researcher who takes, as given, model M and must choose an experiment $\mathscr{D}$ that tests or classifies M. Alternatively, we could imagine a researcher (or referee) who takes, as given, an existing experiment $\mathscr{D}$ and wants to know which models it tests or classifies. In that case, take the labeled permutohedron and "cut" any edge whose label contains a menu from $\mathscr{D}$. The result will be the experiment partition $R_{\mathscr{D}}$, described above. This partition can be thought of as a complete model. Following the discussion that precedes Definition 3, $\mathscr{D}$ will classify any complete model that is a coarsening of $R_{\mathscr{D}}$, and will test any model M such that M_0 equals the union of some set of types in $R_{\mathscr{D}}$.[16]

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 is 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. There are no differentiated pairs that are 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 adjacent to a ranking in M_0. In doing so, it is possible we completely remove the shortest paths between two rankings P and P'. As we show in Proposition 2 in Section VII, two rankings are separated by an experiment if and only if the experiment contains a menu listed along the edges of any 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

[16] One could then define an ordering over experiments based on how fine are the partitions they generate. Certainly if $\mathscr{D}$ is a refinement of $\mathscr{D}'$ then $R_{\mathscr{D}}$ will be a refinement of $R_{\mathscr{D}'}$. It would be interesting to explore further results along these lines.

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 $P'' \in M$ that occurs along any shortest path between P and P' in the full permutohedron $(\mathscr{P}, \mathscr{E})$. 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: $\tilde{L}(E) = \{S \subseteq X : \mathrm{dom}_P(S) \neq \mathrm{dom}_{P'}(S)\}$. That is, the edges are labeled with all the sets for which the neighboring rankings choose differently.

For instance, consider a model where $M_0 = \{adcb, dacb\}$. Its restricted permutohedron is shown in Figure VII. The 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.

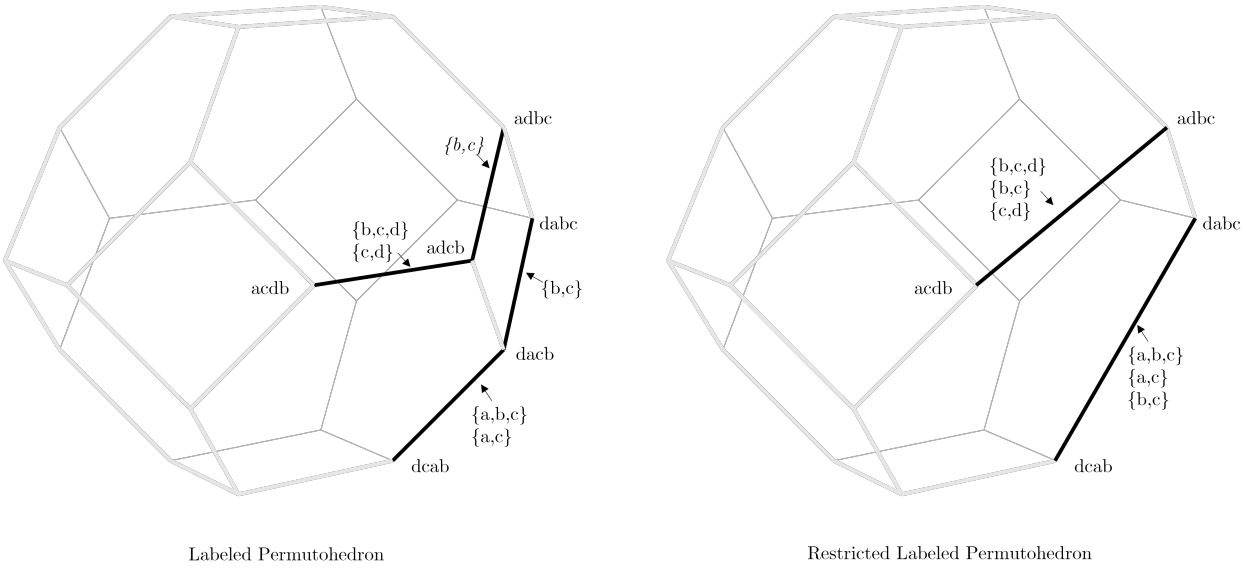

FIGURE VII. The restricted labeled permutohedron for 4 objects $X = \{a, b, c, d\}$ with $M_0 = \{adcb, dacb\}$. (Only the bold edges have been labeled.)

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 on the set $X = \{a,b,c,d\}$. The first type consists of all rankings with a ranked first. The second type is the single ranking $\{bcda\}$. M_0 consists of all remaining rankings. After deleting all vertices in M_0 and their adjacent edges, every shortest path (on the full permutohedron) between each of the $a-$first rankings and $bcda$ is removed. Thus, each of the $a-$first rankings becomes a restricted neighbor of $bcda$.

Now consider the shortest paths on the restricted permutohedron between $abcd$ and $adcb$. Any path between this pair that remains inside the experiment set containing the a-first rankings involves three edges. The shortest path on the restricted permutohedron is a two-edge path passing outside of that experiment set through the vertex $bcda$. The experiment is not convex with respect to the shortest paths on the *restricted* permutohedron. This example is depicted in Figure VIII.

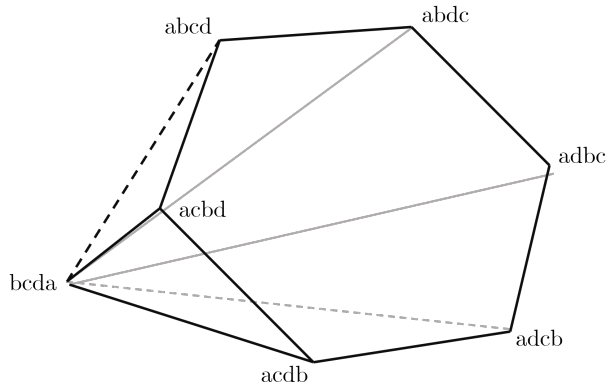

FIGURE VIII. The restricted permutohedron for objects $X = \{a,b,c,d\}$ where t_1 is the set of all $a-$first rankings and $t_2 = \{bcda\}$. Dotted lines show the shortest path between $abcd$ and $adcb$.

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, we only required that the experiment partition is a set of connected subgraphs. We prove this weaker condition within the proof of Theorem 2, although we note that the convexity of the experiment partition on the full permutohedron still plays a key role in this proof.

Finally, recall that Theorem 1 showed that an experiment classifies a complete model if and only if the experiment partition $R_{\mathscr{D}}$ is a refinement of $M = (t_1,\ldots,t_n)$, meaning $r(P) \subseteq t(P)$ for every $P \in \mathscr{P}$. With a restricted model, the requirement is weaker: We can achieve classification even when $r(P) \subseteq t(P) \cup M_0$ because the experimenter classifies types assuming orderings in M_0 are "impossible." All that is required is that $r(P)$ does not include rankings from two different types in the model.

To capture this looser requirement, we can restrict the experiment partition to $\mathscr{P} \setminus M_0$ in the following way: Start with the original experiment partition $R_{\mathscr{D}} = \left(r_1, \ldots, r_q\right)$. For any set $M_0 \subseteq \mathscr{P}$ define the M_0-restricted experiment partition $\tilde{R}_{\mathscr{D}}(M_0) = \left(\tilde{r}_1, \ldots, \tilde{r}_q\right)$ to be the partition of $\mathscr{P} \setminus M_0$ such that $\tilde{r}_i = r_i \setminus M_0$ for each $i \in \{1, \ldots, q\}$. Let $\tilde{r}(P)$ be the partition element containing P. It is straightforward to see that classifying a restricted model $M = (t_1, \ldots, t_n, M_0)$ is equivalent to requiring that $\tilde{R}_{\mathscr{D}}(M_0)$ is a refinement of $(t_1, \ldots, t_n)$, so that $\tilde{r}(P) \subseteq t(P)$ for every P.

We are now ready to state and prove a generalization of Theorem 1 that allows for the classification of restricted models; when $M_0 = \emptyset$ Theorem 2 is identical to Theorem 1.

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*)**.** The following are equivalent:

(1) Experiment $\mathscr{D}$ classifies a model $M = (t_1, \ldots, t_n, M_0)$,
(2) $\mathscr{D}$ separates every restricted boundary pair for model M, and
(3) The M_0-restricted experiment partition $\tilde{R}_{\mathscr{D}}(M_0)$ is a refinement of the partition $(t_1, \ldots, t_n)$.

Proof of Theorem 2. As in Theorem 1, equivalence between (1) and (3) is straightforward, so we focus on the equivalence between (1) and (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.

Before proceeding, we first prove that the partition elements in $\tilde{R}_{\mathscr{D}}(M_0)$ are connected subgraphs.

Lemma 5 ($\tilde{R}_{\mathscr{D}}$ *is a Set of Connected Subgraphs*)**.** Each set $\tilde{r}_i$ in $\tilde{R}_{\mathscr{D}}(M_0)$ 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' have 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 $\tilde{r}(P) = \tilde{r}(P')$.

By Lemma 5, there is a path from P to P' entirely in $\tilde{r}(P)$. 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 $\tilde{r}(P)$, so $\tilde{r}(\hat{P}) = \tilde{r}(\hat{P}')$. 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 using 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).

Choose-k menus expand the set of experiments that test/classify a model. For example, consider objects $X = \{a, b, c\}$ and the complete model in which every ordering is of a separate type. Any experiment that classifies this model must include $D_1 = \{a, b\}, D_2 = \{a, c\}, D_3 = \{b, c\}$. However, if choose-two menus are permitted, the complete model can be classified by asking subjects to choose their *two* favorite objects from $\{a, b, c\}$ and their *single* favorite object from $\{a, b, c\}$.

As another example, the convex preferences model in Section II partitioned the set $\{a, b, c\}$ into $t_1 = \{abc, acb, bac, cab\}, M_0 = \{bca, cba\}$. Recall that with choose-one menus, any experiment that tests the model requires including both $D_1 = \{a, b\}$ and $D_2 = \{a, c\}$. However, if we allow choose-two menus, it is possible to test the model by having subjects choose their two favorite objects from $\{a, b, c\}$.

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-two menus included is shown in Figure IX.

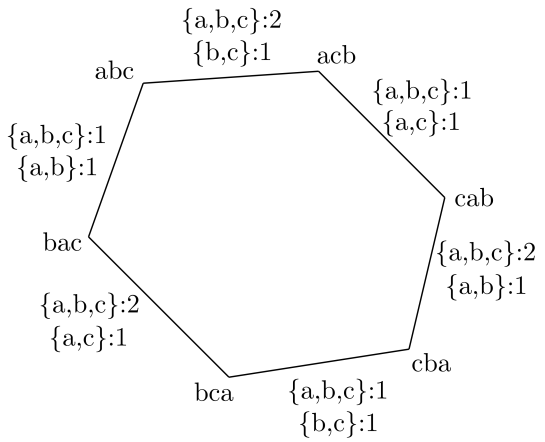

FIGURE IX. The labeled permutohedron for objects $X = \{a, b, c\}$ with choose-two menus included.

In Appendix A, 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 theorems' 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 the ranking P'' chooses differently. Since P'' chooses differently, there must be some pair of 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.

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 labeled permutohedron in studying preferences.

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

Take, for example, the rankings $P = abcd$ and $P' = cadb$. These differ by three transpositions: $T(P,P') = \big\{\{a,c\},\{b,d\},\{c,b\}\big\}$. 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 X. 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 $\{a,c\},\{b,d\},\{c,b\},\{a,b,c\},\{a,c,d\},\{b,c,d\},\{a,b,c,d\}$. The two rankings choose differently from each set. For instance, $abcd$ chooses b from $\{c,b\}$ while $cadb$ chooses c. Furthermore, there is *no other set* for which these two rankings choose differently.

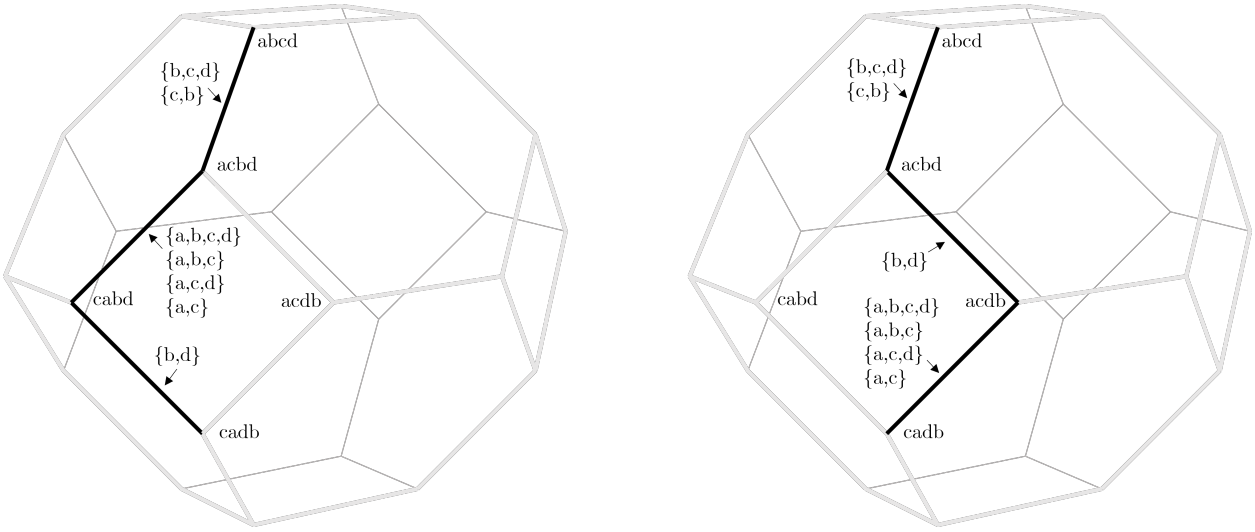

FIGURE X. The Two Shortest Paths from $abcd$ to $cadb$

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. For sufficiency, suppose $\mathscr{D}$ separates P from P', meaning there is some $D'_i \in \mathscr{D}$ such that $\text{dom}_P(D'_i) \neq \text{dom}_{P'}(D'_i)$. And, by way of contradiction, suppose that for every shortest path W from P to P' we have that $\mathscr{D} \cap L(W) = \emptyset$, meaning no set from the experiment appears along the path W. Pick any shortest path $W = (P_1, \ldots, P_v)$ with $P = P_1$ and $P_v = P'$. Then for every P_j on W (with $j < v$) we have that $D'_i \notin L(\{P_j, P_{j+1}\})$, meaning $\text{dom}_{P_j}(D_i) = \text{dom}_{P_{j+1}}(D_i)$. Since this is true for all $j < v$ we conclude that $\text{dom}_P(D'_i) = \text{dom}_{P'}(D'_i)$, a contradiction.

Conversely, let $W = (P_1, \ldots, P_v)$ be a shortest path from P to P' (so $P = P_1$ and $P_v = P'$) and suppose there is some $D'_i \in L(W) \cap \mathscr{D}$, but $\mathscr{D}$ does not separate P from P', meaning for every $D_j \in \mathscr{D}$ we have $\text{dom}_P(D_j) = \text{dom}_{P'}(D_j)$. To simplify notation, for each $1 \leq j \leq v$ define $x_j = \text{dom}_{P_j}(D'_i)$. Since $D'_i \in \mathscr{D}$ and $\mathscr{D}$ does not separate P and P', it must be that $x_1 = x_v$. Now let P_k be the first ordering in $W = (P_1, \ldots, P_v)$ such that $x_k \neq x_{k+1}$; there must exist at least one such ordering since $D'_i \in L(W)$. And since P_k is the first such ordering, $x_1 = x_k$ and so $x_k = x_v$, yet $x_k \neq x_{k+1}$. Since $x_k \neq x_{k+1}$ we have that $\{x_k, x_{k+1}\} \in T(P_k, P_{k+1})$ (meaning P_k and P_{k+1} rank x_k and x_{k+1} differently), which means $\{x_k, x_{k+1}\} \in L(W)$. Also, $\{x_k, x_{k+1}\} \notin T(P, P')$ because both P and P' select x_k (over x_{k+1}) from D'_i. But then the second part of Lemma 3 gives $\{x_k, x_{k+1}\} \notin L(W)$, which is a contradiction. □

Proposition 3 (*All Shortest Paths 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 $W' = (P_1, ..., P_v)$. For all $i < v$, $dom_{P_i}(D) = dom_{P_{i+1}}(D)$. 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. Using the Theorems to Find Minimal Experiments

Suppose an experimenter wants to choose an experiment that is optimal in some well-defined sense. For example, they want to test model M using the fewest number of questions or at the lowest cost. Searching over the set of all possible experiments is obviously wasteful, as the number of possible experiments explodes double-exponentially. Specifically, with $|X|$ objects, the number of possible non-empty and non-singleton menus is $2^{|X|} - 1 - |X|$ and

so the number of non-empty experiments is thus $2^{2^{|X|}-1-|X|}-1$.[17] For $|X|=3$ there are 15 possible experiments, while for $|X|=6$ there are over 144 quadrillion. However, our results characterize the exact subset of experiments that test or classify a given model. This provides a "budget set" of experiments that one might consider. And then, given some objective such as minimizing cost, one can search for the optimal experiment from within that budget set. In this section, we formalize that process.

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 $\mathscr{D}'$. For example, $\mathscr{D}$ may be less costly or involve fewer decisions than $\mathscr{D}'$. An experiment $\mathscr{D}$ is *minimal for testing* M if $\mathscr{D}$ tests M and there is no $\mathscr{D}'$ testing 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.

One interesting objective is finding an experiment that requires the fewest choices. This can be formalized through the *(lexicographic) size ordering*: Let $\mathscr{D}' > \mathscr{D}$ if (1) $\mathscr{D}'$ contains more menus than $\mathscr{D}$, denoted $|\mathscr{D}'| > |\mathscr{D}|$, or (2) $|\mathscr{D}'| = |\mathscr{D}|$ and $\sum_{D' \in \mathscr{D}'} |D'| > \sum_{D \in \mathscr{D}} |D|$.[18] Throughout this section, when we refer to a "minimal experiment", we mean with respect to the lexicographic size order, but we emphasize that our general results are not specific to a particular experiment ordering.

In this section, we first describe an algorithm for finding minimal experiments. We then show how minimal experiments can be used to help cluster subjects into economically-meaningful types. Following this, we provide two examples of how minimal experiments can be used to generate and study elicitation methodology in experimental economics.

An Algorithm for Finding Experiments of Minimal Size

We now show how identifying the minimal experiment under the lexicographic size order can be solved as a straightforward integer binary linear program. Extending this idea to other experiment orderings is straightforward.

The algorithm can be understood as consisting of two major 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

[17] There are $2^{|X|}$ subsets, minus the empty set and the $|X|$ singleton sets. So there are $2^{2^{|X|}-1-|X|}$ sets of subsets. We remove the empty experiment to get $2^{2^{|X|}-1-|X|}-1$.

[18] This particular ordering is complete but not total. There may be multiple minimal experiments that achieve some objective. When including choose-k menus, choosing the experiment ordering is not as straightforward when the goal is to minimize the number of subject choices. For instance, it is not obvious whether the single-menu experiment $D_1 = \{a,b,c\} : 2$ is larger, smaller, or equal to the experiment with $D_1 = \{a,b\} : 1$ and $D_2 = \{a,c\} : 1$.

been enumerated, the algorithm proceeds to solve the resulting set cover problem by converting it into a linear program. This part is identical whether or not the model is restricted.

Algorithm:

1. Input a model M over objects X.
2. Is the model complete or not?
   2.a. *Complete Model (or Testing a Model)*: For each pair of rankings P and P' in different sets in M, find the transpositions $T(P,P')$. If the $|T(P,P')| = 1$, rankings are a boundary pair.
   2.b. *Restricted Model*: 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.
3. For each boundary pair $\{P,P'\}$, determine the sets in $L(\{P,P'\})$ for which P and P' choose differently.[19]
4. Let $E = (E_1,...,E_m)$ be the set of m boundary pairs and $S = \{S_1,...,S_l\}$ be the union of the sets appearing on the edges of those boundary pairs (note that $|S| = l$). Construct an $m \times l$ matrix O such that $O_{(i,j)} = 1$ if set S_j appears on the edge for boundary pair i, and $O_{(i,j)} = 0$ otherwise.
5. Construct a lexicographic cost vector c of length l where $c_j = 1 + \frac{|S_j|}{|X|\cdot l}$.[20] (Different cost functions can be used here to represent different experiment orderings.)
6. Solve the resulting set cover problem by integer binary linear programming. Below, vectors are column vectors, $\mathbb{1}_m$ is the length-m vector of ones, and c^T is the transpose of c.
$$\begin{aligned} &\text{Minimize} \quad c^T \cdot x \\ &\text{subject to} \quad O \cdot x \geq \mathbb{1}_m \;\; and \quad x \in \{0,1\}^l \end{aligned}$$
7. Each solution x^* defines a minimal experiment, wherein the minimal experiment includes S_j if and only if $x^*_j = 1$. The constraint $O \cdot x \geq \mathbb{1}_m$ ensures that at least one menu from every edge between boundary pairs is included in the experiment.

Example 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

[19]Recall from above the simple characterization of these sets: If $T(P,P') = \{\{x,x'\}\}$ then define $B(\{x,x'\};P) = \{x,x'\} \cup \{z \in X : (\forall y \in \{x,x'\})\ yPz\}$. Then $D_i \in L(\{P,P'\})$ if and only if (1) D_i contains $\{x,x'\}$, and (2) $D_i \subseteq B(\{x,x'\},P)$. Thus, $|L(\{P,P'\})| = 2^{|B(\{x,x'\},P)|-2}$.

[20]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 $|X|\cdot l$ since $|X|$ is the number of objects and l is the total number of sets appearing on the edges between boundary pairs. Thus, the weight $\frac{1}{|X|*l}$ ensures the costs are lexicographic, prioritizing the number of sets over set size.

each boundary pair, respectively, are $\{\{a,b,c\},\{a,b\}\}$, $\{\{a,c\}\}$, $\{\{a,b,c\},\{a,c\}\}$, and $\{\{a,b\}\}$. 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:

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

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

We do not claim that this algorithm is particularly efficient or can scale to find minimal experiments of arbitrary complexity. While our boundary pair result drastically reduces the complexity of finding a minimal experiment relative to brute force search, both finding the boundary pairs for a model and solving the resulting set cover problem are computationally difficult.[21]

In Appendix B we provide results of simulations for finding minimal experiments for randomly generated models with numbers of objects ranging from $|X| = 4$ to 10 and the number of types (each with three rankings) ranging from 2 to 7. We ran 100 models for each combination of these parameters and, for each random model, found the minimal experiment for classifying only as well as testing and classifying. Runtime grows exponentially with $|X|$. Although runtime also grows with the complexity of the model (number of types), $|X|$ is the dominant driver.

Despite exponential growth, the algorithm remains practically useful for moderately sized models where manual search or brute force would be difficult or impossible. Most trials were completed in seconds.[22] We offer the following example which also demonstrates the ability of this algorithm to extract patterns in models that are not easily seen otherwise.

Example (*Nine-Object Experiment*). Consider the following (randomly generated) restricted model over nine objects with $X = \{a,b,c,d,e,f,g,h,i\}$:

$$t_1 = \{dcafiehbg, cdgeiafhb, ecigbhafd\}$$
$$t_2 = \{gaifbdceh, gdhafcieb, acdbeihgf\}$$
$$t_3 = \{bfaehdigc, dhbegcfai, gfdbichae\}$$

The number of possible experiments for nine objects has 152 digits. Despite this complexity, the algorithm presented above finds a minimal experiment to classify this restricted model in about three seconds. The computer spends roughly equal amounts of time applying

[21] Set cover problems are generically NP-hard (Korte and Vygen, 2008).

[22] Our code is not extensively optimized and ran only on a single core.

the boundary-pair theorem and solving the resulting linear programming problem for set-selection. It has just one menu: $\{a,b,c\}$.[23] To confirm this classifies the model, notice that in t_1 every ranking would choose c. In t_2, every ranking would choose a. In t_3, every ranking would choose b. It is remarkable how quickly the algorithm finds this pattern, though it is obvious in hindsight.

This example demonstrates another way of thinking about minimal experiments. If an experiment classifies a model, the sets of choice profiles for that experiment represent choice patterns that fully describe the model. In this sense, a minimal experiment extracts the smallest set of choice patterns that fully describe a model. In the next section, we leverage this observation and show how minimal experiments can be used to interpret non-parametric preference clusters.

Application: Non-Parametric Preference Clustering

Clustering can be seen as an approach to building a model (determining types) from data. For example, Fehr and Charness (2025) report that social preference data from a broad population of subjects tends to produce three qualitative types: "Altruistic", "Inequality Averse", and "Predominantly Selfish." A clustering such as this can be viewed as a model in our framework. And this cluster-based model can then be used to classify future subjects.

Specifically, we envision a research agenda that proceeds as follows: (1) Have a "training set" of representative subjects participate in an experiment that elicits their entire ranking. This experiment might require the subject to make many choices, though there are ways to incentivize reports of an entire ranking (Bateman et al., 2007). (2) Use an established clustering technique to sort preference orderings into clusters, each of which becomes a type in our framework. We describe one such technique below. A distance metric such as Kendall tau may be used to identify preferences that are "near enough" to a given cluster to be included in that cluster. And the researcher may choose to leave some orderings uncategorized if they are considered too far from any given cluster. The result will be a model M that is complete if all orderings are categorized or restricted if some are not. (3) Use the algorithm above to identify a minimal experiment $\mathscr{D}$ to classify model M. This experiment can be used to provide interpretation of the various clusters, as we demonstrate below. (4) If desired, run experiment $\mathscr{D}$ on a new "testing set" of subjects to estimate the population frequency of each identified type. Using a minimal experiment ensures this estimation is done efficiently

[23] For testing and classifying the model, the minimal experiment is substantially more complex, but the algorithm still produces a 24-menu minimal experiment in about 3.5 seconds: $\{c,d,h\}$,$\{a,c,g\}$,$\{a,f\}$,$\{f,h,i\}$,$\{b,e,i\}$,$\{e,g,i\}$,$\{e,h\}$,$\{b,h\}$,$\{b,g\}$,$\{d,g,i\}$,$\{a,i\}$,$\{b,f,h\}$,$\{c,e,h\}$,$\{c,i\}$,$\{g,i\}$,$\{a,h\}$, $\{d,f\}$,$\{b,d\}$,$\{c,f\}$,$\{b,e\}$,$\{g,h\}$,$\{f,g\}$,$\{a,e\}$,$\{c,g\}$. While complex, we note that the experiment contains 12 fewer menus than the experiment that includes every binary pair, which would require 36 menus.

and does not require entire preference orderings to be revealed. In some contexts, this may also be preferred by subjects concerned about privacy.

In the Fehr and Charness (2025) example above, the clusters each have a clear interpretation, but in many settings they do not (see Bertsimas et al., 2021). One benefit of the minimal-experiment approach is that the choices each type would pick in the minimal experiment can actually provide an economic interpretation of each type. As a very simple example, if the minimal experiment for some three-type clustering consisted of the single menu $D_1 = \{a, b, c\}$ then the clusters would be defined as "those who like a," "those who like b," and "those who like c." Thus, the various clusters can be interpreted simply by the choices that are used to identify them.

Step (2) of the above procedure requires a statistical method to identify clusters in the training data. One common method is to use a "mixture model," where the data is assumed to be generated by a collection of predefined sub-models. Often, each sub-model (or, cluster) is defined by a central ranking and a dispersion parameter that allows for some (probabilistic) deviation from that central ranking (Murphy and Martin, 2003, *e.g.*). There are also non-parametric approaches for clustering, such as *Partitioning Around Medoids* (PAM). PAM begins by selecting n representative points in the dataset, called medoids, as initial cluster "centers." Each data point is then assigned to the cluster corresponding to the nearest medoid according to a chosen distance measure, such as Kendall tau. The algorithm iteratively swaps existing medoids with non-medoids, evaluating whether the swap reduces the total dissimilarity between points and their assigned medoids. This iterative "hill-climbing" optimization continues until no further improvement is possible (see Kaufman and Rousseeuw, 2009, *e.g.*).

Although not often used on preference or choice data, PAM is a compelling option for our purposes. It is a simple method that relies only on a calculated proximity matrix and does not hinge on any distributional assumptions. However, unlike parametric mixture models, this approach does not provide information about the spread or diversity of the rankings within the cluster, making it harder to interpret. PAM represents a cluster only by a single medoid, typically used as a representative for interpretation (Everitt et al., 2011). But there is no guarantee that the characteristics of the medoid accurately reflect the characteristics of the entire cluster. Here, the choices each cluster would pick in the minimal experiment may provide a more intuitive interpretation of the various clusters, as described above.

To demonstrate this approach—and keeping with the breakfast theme established in our introduction—we apply PAM to cluster data from Green and Rao (1972), which includes preferences among seven breakfast items for 42 subjects.[24] We use the Kendall tau distance (the number of transpositions between two rankings) for proximity. It is also the length of

[24]To simplify the data, we removed several highly correlated items such as "hard rolls with butter" and "buttered toast."

any shortest path between two rankings in the permutohedron (see Section VII). For illustrative purposes, we calculate a two-cluster solution. The two medoids identified, labeled as P_1^* and P_2^*, appear in Table I. All rankings observed from the 42 subjects are assigned to the cluster whose medoid is closer, and any ranking not observed is considered outside the model (thus, in M_0).[25]

Rank	$\mathbf{P_1^*}$	$\mathbf{P_2^*}$
1	Coffee Cake	Buttered Toast
2	Blueberry Muffin	English Muffin
3	Cinnamon Toast	Cinnamon Toast
4	Jelly Donut	Coffee Cake
5	English Muffin	Jelly Donut
6	Buttered Toast	Toast Pop-up
7	Toast Pop-up	Blueberry Muffin

TABLE I. The medoids of the PAM clustering of breakfast item preferences from Green and Rao (1972).

Can we use these two medoids to "eyeball" an interpretation of the two clusters? For example, it might be tempting to say that cluster 1 likes blueberry muffins, while cluster 2 does not. But it turns out that this is not an accurate representation of all preferences in each cluster: There is a ranking in cluster 2 that actually ranks blueberry muffins first. Or it may appear that cluster 1 prefers coffee cake while cluster 2 prefers buttered toast, but this is also inaccurate since there is a ranking in cluster 1 that ranks buttered toast higher. Instead, if we find a minimal experiment for classifying these two clusters, then the choices separate the clusters, and so the experiment can be used to help interpret the difference between them.

To illustrate, the minimal (in terms of size) experiment for classifying these two clusters consists of three menus:

$$D_1 = \{\text{Coffee Cake, Buttered Toast}\},$$

$$D_2 = \{\text{Jelly Donut, Blueberry Muffin}\}, \text{ and}$$

$$D_3 = \{\text{Toast Pop-up, Jelly Donut}\}.$$

[25]Alternatively, one could generate a complete model by taking a type to be *all* rankings closest to some medoid whether they appear in the data or not. This would, however, force the resulting minimal experiment to separate potentially "unusual" rankings that would never be observed.

There are eight combinations of choices a subject could make from these three binary menus. And which combination a person chooses perfectly identifies to which cluster they belong. As shown in Table II, four of these choice combinations lead to a subject being classified in cluster 1, three lead to cluster 2, and one leads to a subject being classified as outside the model.

	Choices From			**Number of**
Cluster	**D_1**	**D_2**	**D_3**	**Subjects**
1	Coffee Cake	Blueberry Muffin	Jelly Donut	13
	Coffee Cake	Jelly Donut	Jelly Donut	13
	Coffee Cake	Blueberry Muffin	Toast Pop-up	2
	Buttered Toast	Blueberry Muffin	Jelly Donut	1
2	Buttered Toast	Blueberry Muffin	Toast Pop-up	8
	Buttered Toast	Jelly Donut	Jelly Donut	4
	Buttered Toast	Jelly Donut	Toast Pop-up	1
None	Coffee Cake	Jelly Donut	Toast Pop-up	0

TABLE II. The choice combinations that lead to classifying a subject into each cluster given by the PAM clustering procedure.

Table II reveals that clusters are almost perfectly determined by the choice from D_1; people are either a "coffee cake type" or a "buttered toast type." The one exception (which consisted of a single subject here) is that if someone likes buttered toast, then we need their other two choices to see if they should actually be in cluster 1.

In fact, one could use this minimal experiment to redefine the clusters given by the PAM procedure, moving the one exception from cluster 1 into cluster 2. The new model would have a very simple interpretation: those who like coffee cake and those who like buttered toast. And the minimal experiment for classifying these new clusters is very simple, consisting only of D_1.

Alternatively, the orderings consistent with the one exceptional choice combination could be dropped from the model and moved into M_0 (the "None" category in Table II). If, for example, we remove the 10% of rankings from each cluster that are farthest from their associated medoid, then this combination of choices is removed from the model and the minimal experiment again reduces to D_1. Doing so leaves unchanged the classification of 97.6% of the 42 subjects (all but one) and, therefore, is arguably a small change that significantly simplifies the clusters.

This suggests that the size of the minimal experiment could even be used as a measure of the complexity of a clustering. The original clusters in Table II required three binary-choice menus, while the modified versions require only one.

Our example uses the PAM clustering method, but we emphasize that the same approach could be used to aid interpretation for any clustering method, including more traditional mixture models. In fact, the clusters may come from a researcher's prior over preferences based on past work, and not from a new "training set" of subjects. Regardless of how the clustering is formed, the minimal experiment provides the simplest way to classify data collected in the future into precalculated clusters without collecting complete rankings. This could be beneficial in experimental economics to classify subjects quickly in follow-up studies. It could also be useful in consumer preference analysis to quickly classify new consumers into preference clusters for use in marketing, recommendations, or demand estimation.[26]

Application: Eliciting Ranges of Beliefs

Suppose a researcher wants to elicit a subjective probability p that an event E will occur. And suppose the researcher does not need a precise belief; it is sufficient to categorize the belief into three categories: $p \in [0, 0.4)$, $p \in (0.4, 0.6)$, and $p \in (0.6, 1]$. There are three relevant choice objects available: $l_{0.6}$ is a lottery that pays \$10 with objective probability 0.6, t (for "true") is an act that pays \$10 if event E is true, and f (for "false") is an act that pays \$10 if E is false.

In terms of these objects, 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 ($M_0 = \{t f l_{0.6}, f t l_{0.6}\}$).

To classify the restricted model, we construct the restricted permutohedron.[27] In this case, the restricted permutohedron is simply the graph induced by removing the vertices in M_0 from the permutohedron.[28] The restricted permutohedron for this model is shown in the left panel of Figure XI.

[26] See Müllensiefen et al. (2018), which discusses the use of clustering to understand brand preferences. (Zhang et al., 2016) discusses the use of clustering to improve recommendation systems.

[27] Recall that classifying a restricted model implies the researcher assumes preferences in M_0 are impossible. Here, rankings in M_0 are impossible under the assumption that each individual has a subjective probability $p \in [0, 1]$ for event E, $1 - p$ for its complement, and orders all bets by their probability of \$10.

[28] Recall from Section V there are some models where the construction of restricted permutohedra requires creating new edges. However, here, the remaining vertices are still connected after deleting all the vertices in M_0 and their associated edges.

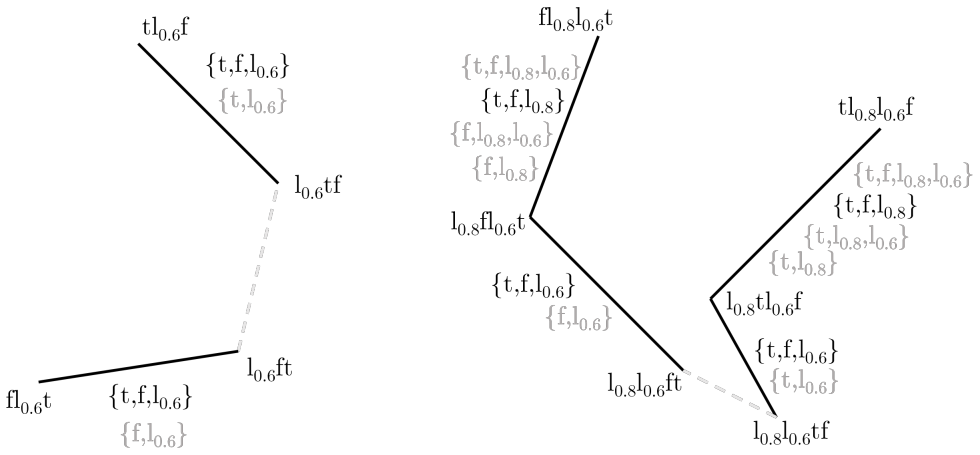

FIGURE XI. The restricted permutohedra for three-category (left) and five-category (right) belief elicitation. Only the edges between boundary pairs (shown in bold) have been labeled. Menus used in the minimal experiment are shown in bold.

The minimal experiment (in terms of size) for the three-category belief elicitation involves just one set: $D_1 = \{t, f, l_{0.6}\}$. This set appears on both 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 restricted model:[29]

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

The restricted permutohedron for this model, shown in the right panel of Figure XI, 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 $D_2 = \{t, f, l_{0.8}\}$.[30] The experiment might appear this way:

[29]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.

[30]The algorithm for finding minimal experiments presented later in this section reveals that a minimal experiment for 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 $D_4 = \{l_{0.6}, t\}$. Notice that D_2 tests whether $l_{0.8} \succ l_{0.6}$ and D_3 and D_4 jointly test whether $l_{0.6}$ is preferred to at least one of t and f while also aiding in classification.

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

It is also possible to find the minimal experiment for belief elicitation with a larger number of categories. 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*. Healy and Leo (2025) expands on the theory of ternary price lists and demonstrates that they have theoretical properties that improve on the two existing methodologies for belief elicitation: binary price lists and binarized scoring rules. We believe this example demonstrates the practical value of our results to help generate novel and useful experimental methodologies.

Application: Minimality of Price Lists

Multiple price lists (MPLs) present subjects with a structured series of binary choices, typically in a table format, and are used to bound indifference points in a preference relation. Many MPLs have a simple structure. In each binary choice, there is one fixed object and one object that changes across the rows/choices. Such MPLs have been used in a variety of contexts, including in eliciting willingness-to-pay, risk preferences, and time preferences (Holt and Laury, 2002; Andersen et al., 2006). In this section, we demonstrate that the ubiquity of price lists might be partially due to the fact that they are minimal (in terms of size)—even among extended experiments—for a certain class of restricted models.

Let $A = \{a_1, ..., a_{|A|}\}$ be a set of objects (such as dollar amounts) with a natural ordering such that it is reasonable to assume $a_i \succ a_j$ if and only if $i < j$. Consider the goal of learning about preferences on the set $X = A \times \{x\}$, where x is the object of interest and the researcher wants to know how x ranks among the ordered items in A. For example, if A consists of dollar amounts and x is a cheeseburger, finding the ranking of x among A gives an estimate of the subject's willingness to pay for the cheeseburger. Let P_i be the order where x is ranked below $i+1$ elements of A so that:

$$P_1 = xa_1...a_{|A|}, \quad P_i = a_1...a_{i-1}xa_i...a_{|A|}, \quad P_{|A|+1} = a_1...a_{|A|}x$$

This goal can be written as the following restricted model with types $t_1 = \{P_1\}, ..., t_i = \{P_i\}, ..., t_{|A|+1} = \{P_{|A|+1}\}$. We refer to such a model as a *linear preference model*. The restricted

permutohedron for a linear preference model is a path (linear) graph that is a convex set in the full permutohedron of these objects. Thus, the shortest path between each of the rankings remains intact in the restricted permutohedron. Every pair of the form (P_i, P_{i+1}) is a restricted boundary pair, since the rankings differ by a single adjacent transposition and every ranking is in a singleton set in the restricted model.

Allowing extended experiments (including choose-k menus), the set of menus on the edge of the boundary pair (P_i, P_{i+1}) consists of all menus of the form $\{x, a_i\} \bigcup \tilde{A} : k$ where $\tilde{A} \subseteq \{a_1, ..., a_{|A|}\} \setminus a_i$ and there are $k-1$ elements $a \in \tilde{A}$ such that $a \succ a_i$. All such menus are such that the kth choice of P_i is a_i while the kth choice of P_{i+1} is x. For example, $\{x, a_i\} : 1$ and $\{x, a_i, a_{i-1}\} : 2$ are on the edge as long as $i \geq 2$.

A consequence is that the menus on the various edges of the restricted permutohedron (which are all between boundary pairs) are disjoint. Since the experiment must contain a menu from every edge between boundary pairs, the experiment must contain at least $|A|$ menus. To find the minimal experiment in terms of size, choose the smallest menu on each edge, which is $\{x, a_i\} : 1$. This yields an experiment with menus $D_i = \{\{x, a_i\} : 1\}_{i=1}^{|A|}$, which is exactly a multiple price list (MPL).

IX. Discussion

Throughout the paper, we assume that preferences and choices are deterministic. One could imagine that subjects instead make stochastic or "noisy" choices. As mentioned in the Introduction, our framework cannot be used to study stochastic choice for a variety of reasons. However, a direction one could take is an econometric estimation (rather than exact identification) of stochastic choice functions. This likely would require larger-than-minimal experiments, compared to the deterministic case. While considering such a framework would certainly be useful, it deviates too far from the structure assumed here and is therefore left for future work.

In relation to this problem, researchers may be able to leverage redundant sets to determine *if* subjects are choosing stochastically. Whenever there are two or more distinct sets on an edge between boundary pairs, multiple sets can be included in the experiment to "cover" that edge. Since only one set is required for each boundary pair, by including multiple sets the data becomes redundant and can classify subjects even if choices from one of the redundant sets are removed from the data. This allows researchers to classify subjects using distinct sets of choices. If a subject is classified into different types using different data sets generated in this way, this refutes the assumption of deterministic preferences and therefore provides strong evidence that they are choosing stochastically.

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 cannot. 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. However, there is typically no requirement that the data be rich enough to guarantee that the test will be conclusive. Our approach instead searches for choice menus from X such that the resulting data will always be rich enough to guarantee a conclusive test.

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.[31] 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.[32]

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 $D_1 = \{a,d\}$, $D_2 = \{a,b\}$, $D_3 = \{b,c\}$. 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.[33] 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 that the experiment be designed so that the test is always conclusive. 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

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

[32] If X does not contain all possible choice objects then of course our approach may also fail to detect violations of the model if they occur outside of X.

[33] Specifically, if $d = \alpha_1 a + \alpha_2 b + \alpha_3 c$ then $\lambda_1 = 1/(\alpha_1 + 2\alpha_2 + 3\alpha_3)$, $\lambda_2 = (\alpha_2 + \alpha_3)/(\alpha_1 + 2\alpha_2 + 3\alpha_3)$, $\lambda_3 = \alpha_3/(\alpha_1 + 2\alpha_2 + 3\alpha_3)$.

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.

There could also be settings where a researcher wants to add constraints on which types of menus are used in their experiment. For example, they may want to restrict their experiment to contain only binary choice menus. The permutohedron approach could still be used, and the constraint on admissible menus would be added when searching for a set of menus that covers the set of boundary pairs for the model. For example, in Part 2 of our algorithm, let $O_{(i,j)} = 1$ if set S_j appears on boundary pair i *and* satisfies the new constraint. Under some constraints a solution may no longer exist, though if the constraint admits at least all of the binary menus, then a solution must exist since the entire preference relation can be elicited via binary menus.

Another benefit of our approach is that models are defined very generally and can be applied at different "levels" of analysis. For example, consider one researcher interested in classifying subjects according to the rank-dependent expected utility model (RDEU; Quiggin (1982)). For them, the relevant model would have types that represent different probability weighting functions or different levels of risk aversion. Another researcher might instead be interested in testing the comonotonic sure-thing principle, which is a fundamental axiom of RDEU. For this researcher the model would have two types: those preferences orderings that satisfy the axiom, and those that do not. Thus, any sort of testing or classifying exercise can fit into our framework by defining the model appropriately for the given problem.

One limitation of our method is it takes as given the set of alternatives X.[34] 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, 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' does not 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 relatively small. Thus,

[34]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 $X = \bigcup_{S_j \in \mathscr{D}} S_j$.

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.

Another interesting open question is what can be said if we relax the assumption of transitivity. In that case, we might replace preference orderings with choice functions that identify a choice for each possible menu D. For example, with $X = \{a, b, c\}$ there are 24 possible choice functions. Then we can define a graph similar to a permutohedron with choice functions as vertices, where two choice functions are neighbors if they differ only in their choice in one menu. For $X = \{a, b, c\}$ this graph would have 24 vertices each with 5 edges, for a total of 60 edges. Each edge is then labeled with the (unique) menu that separates those two choice functions. Any model would then be a partition of this graph, just as in the case of permutohedra with transitivity. We conjecture that our results would go through: An experiment would classify a complete model if and only if it separates every boundary pair on this graph. Unfortunately, this approach requires greater computation to construct the relevant graphs, since there are many more nodes and edges.[35]

We also conjecture that if we consider any choice-function model that assumes transitivity and nothing more (meaning M_0 contains exactly those choice functions that violate transitivity) then we can construct the "restricted choice-function graph" following the procedure from Section V and the result will be the original permutohedron. Thus, transitivity itself can be viewed as a model in this larger framework, and our methods can be used to derive the permutohedron as the relevant graph for that case.[36]

Our approach uses an "all-or-nothing" approach to testing a model, requiring that subjects be perfectly classified into one of several preference types. But what if the researcher has a prior over which types (or which preferences) are most likely and is mainly interested in distinguishing those that are more likely? This can be accommodated in two ways: First, the researcher can simply coarsen the model by either combining or removing unlikely types. The resulting minimal experiment would necessarily shrink as a result. Second, the researcher can generate the minimal experiment for the original model, but then remove

[35]For complete models under the choice function approach, the set-selection step of finding a minimal experiment becomes trivial since each edge contains only one menu. In this sense, comparing the overall computational requirements between the preference ordering and choice function approaches will depend on the particular model being tested or classified.

[36]Similarly, it may be possible to generalize the labeled permutohedron to include indifference by constructing a graph with weak orderings as the vertices. However, the weak incentive compatibility problem would cause issues for the real-world reliability of type identification.

menus from that experiment which they believe are unlikely to be consequential. The (M_0-restricted) experiment partition from this simplified experiment would then constitute a new model, and it may be coarser than the original.

In Section VIII, we focus 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 on the basis of 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.

REFERENCES

Alonso, S., Chiclana, F., Herrera, F., Herrera-Viedma, E., Alcalá-Fdez, J., and Porcel, C. (2008). A consistency-based procedure to estimate missing pairwise preference values. *International Journal of Intelligent Systems*, 23(2):155–175.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9(4):383–405.

Astrachan, O. (2003). Bubble sort: an archaeological algorithmic analysis. *SIGCSE Bull.*, 35(1):1–5.

Atwood, C. L. (1969). Optimal and Efficient Designs of Experiments. *The Annals of Mathematical Statistics*, 40(5):1570–1602.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in Experiments: A Theoretical Analysis. *Journal of Political Economy*, 126(4):1472–1503.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2020). Incentives in experiments with objective lotteries. *Experimental Economics*, 23(1):1–29.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2021). Constrained preference elicitation. *Theoretical Economics*, 16(2):507–538.

Bateman, I., Day, B., Loomes, G., and Sugden, R. (2007). Can ranking techniques elicit robust values? *Journal of Risk & Uncertainty*, 34:49–66.

Berge (1971). *Principles of Combinatorics*. Academic Press.

Bertsimas, D., Orfanoudaki, A., and Wiberg, H. (2021). Interpretable clustering: an optimization approach. *Machine Learning*, 110(1):89–138.

Chiclana, F., Herrera-Viedma, E., and Alonso, S. (2009). A Note on Two Methods for Estimating Missing Pairwise Preference Values. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6):1628–1633.

Everitt, B., Landau, S., Leese, M., Stahl, D., and Safari, a. O. M. C. (2011). *Cluster analysis, 5th edition*. John Wiley & Sons.

Fehr, E. and Charness, G. (2025). Social preferences: fundamental characteristics and economic consequences. *Journal of Economic Literature*, 63(2):440–514.

Fishburn, P. C. (1975). Separation theorems and expected utilities. *Journal of Economic Theory*, 11(1):16–34.

Gaiha, P. and Gupta, S. K. (1977). Adjacent Vertices on a Permutohedron. *SIAM Journal on Applied Mathematics*, 32(2):323–327.

Green, P. E. and Rao, V. R. (1971). Conjoint Measurement- for Quantifying Judgmental Data. *Journal of Marketing Research*, 8(3):355–363.

Green, P. E. and Rao, V. R. (1972). Applied multidimensional scaling: A comparison of approaches and algorithms. *(No Title)*.

Green, P. E. and Srinivasan, V. (1978). Conjoint Analysis in Consumer Research: Issues and Outlook. *Journal of Consumer Research*, 5(2):103–123.

Guilbaud, G. T. and Rosenstiehl, P. (1963). Analyse algébrique d'un scrutin. *Mathématiques et sciences humaines*, 4:9–33.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1934). *Inequalities*. Cambridge University Press.

Healy, P. J. and Leo, G. (2025). Ternary Price Lists for Belief Elicitation. Working Paper.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92(5):1644–1655.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kemeny, J. G. (1959). Mathematics without Numbers. *Daedalus*, 88(4):577–591.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Kendall, M. G. (1948). *Rank Correlation Methods*. C. Griffin.

Kiefer, J. (1959). Optimum Experimental Designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304.

Korte, B. and Vygen, J. (2008). *Combinatorial optimization: theory and algorithms*. Springer.

Lambert, N. S. (2019). Elicitation and Evaluation of Statistical Forecasts. Working Paper.

Luce, R. D. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1):1–27.

Murphy, T. B. and Martin, D. (2003). Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, 41(3-4):645–655.

Müllensiefen, D., Hennig, C., and Howells, H. (2018). Using clustering of rankings to explain brand preferences with personality and socio-demographic variables. *Journal of Applied Statistics*, 45(6):1009–1029.

Pukelsheim, F. (2006). *Optimal design of experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia.

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

Rousseas, S. W. and Hart, A. G. (1951). Experimental Verification of a Composite Indifference Map. *Journal of Political Economy*, 59(4):288–318.

Schulman, R. S. (1979). A Geometric Model of Rank Correlation. *The American Statistician*, 33(2):77–80.

Smucker, B., Krzywinski, M., and Altman, N. (2018). Optimal experimental design. *Nature Methods*, 15(8):559–560.

Yu, P. L. H., Gu, J., and Xu, H. (2019). Analysis of ranking data. *WIREs Computational Statistics*, 11(6):e1483.

Zhang, J., Lin, Y., Lin, M., and Liu, J. (2016). An effective collaborative filtering algorithm based on user preference clustering. *Applied Intelligence*, 45(2):230–240.

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* is a family of tuples $\mathscr{D}^e = \{(D_1,k_1),(D_2,k_2),...,(D_n,k_n)\}$. Typical elements (D_i,k_i) consist of a menu $D_i \subseteq X$ with $|D_i| \geq 2$ and number of choices $k_i < |D_i|$. The interpretation is that each D_i is a menu from which the subject must choose their top k_i most-preferred elements. We define the following choice function:

$$\text{dom}_P^k(X') = \{C \subseteq X' : |C| = k, \forall (x,y) \in C \times (X' \setminus C) : xPy\}.$$

Since all orders are assumed to be antisymmetric, $\text{dom}_P^k(X')$ is unique. 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 $(D_i,k_i) \in \mathscr{D}^e$ such that $\text{dom}_P^k(D_i) \neq \text{dom}_{P'}^k(D_i)$.

Our definitions of the experiment partition $R_{\mathscr{D}^e}$, 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 $(D_i,k_i) \in \mathscr{D}^e$ for which $C = dom_P^{k_i}(D_i) \neq dom_{P''}^{k_i}(D_i) = C''$. However, since $r(P) = r(P')$, $dom_P^{k_i}(D_i) = dom_{P'}^{k_i}(D_i) = C$. Since $C \neq C''$ there is some $x \in C$, $x'' \notin C$, $x'' \in C''$, $x \notin C''$. Thus, for P and P', xPx'', $xP'x''$ and for P'', $x''P''x$. Thus, x and x'' must be inverted at least twice on the path W and so the set $\{x,x''\}$ appears at least twice 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.

We begin by extending Lemma 4 to extended experiments.

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 $t_i \in M$

Proof. 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. □

Theorem 3 (*Generalization of Theorem 1 to Extended Experiments*)**.** The following are equivalent:

(1) Experiment $\mathscr{D}^e$ classifies a complete model $M = (t_1, \ldots, t_n)$,
(2) $\mathscr{D}^e$ separates every boundary pair for model M, and
(3) The experiment partition $R_{\mathscr{D}^e}$ is a refinement of the model partition M.

Proof of Theorem 1. Equivalence between (1) and (3) follows immediately from definitions, so we focus on proving that (1) if and only if (2).

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 separated.

For sufficiency, we will use Lemma 6 to 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 Proposition 4. The entire proof is included here for completeness.

For any set $M_0 \subseteq \mathscr{P}$ define the M_0-restricted experiment partition $\tilde{R}_{\mathscr{D}^e}(M_0) = (\tilde{r}_1, \ldots, \tilde{r}_q)$ to be the partition of $\mathscr{P} \setminus M_0$ such that $\tilde{r}_i = r_i \setminus M_0$ for each $i \in \{1, \ldots, q\}$. Let $\tilde{r}(P)$ be the partition element containing P.

Theorem 4 (*Generalization of Theorem 2 to Extended Experiments*)**.** The following are equivalent:

(1) Experiment $\mathscr{D}^e$ classifies a model $M = (t_1, \ldots, t_n, M_0)$,
(2) $\mathscr{D}^e$ separates every restricted boundary pair for model M, and

(3) The M_0-restricted experiment partition $\tilde{R}_{\mathscr{D}^e}(M_0)$ is a refinement of the partition $(t_1,\ldots,t_n)$.

Proof of Theorem 4. As in Theorem 3, equivalence between (1) and (3) is straightforward, so we focus on the equivalence between (1) and (2).

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 separated.

Before proceeding to prove sufficiency, 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' are 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 full 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 $\tilde{r}(P)$. 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 $\tilde{r}(P)$, so $\tilde{r}(\hat{P}) = \tilde{r}(\hat{P}')$. Thus, we have a boundary pair that is not separated, completing the proof. □

APPENDIX B. COMPUTATION TIME

To address the question of computational complexity, we run the algorithm described in on a randomly generated models of different sizes. Specifically, we vary the number of objects from $|X| = 4$ to 10 and vary the number of types in the model from $n = 2$ to 7. Each type consists of three randomly chosen rankings. We run the algorithm for classifying the (restricted) model, and again for classifying and testing the model. We run each 100 times and report the mean and standard deviation of the computation time (in seconds) in Table III.

n types	test model?	$\|X\| = 4$	$\|X\| = 5$	$\|X\| = 6$	$\|X\| = 7$	$\|X\| = 8$	$\|X\| = 9$	$\|X\| = 10$
2	FALSE	0.02 (0.01)	0.06 (0.01)	0.13 (0.02)	0.31 (0.06)	0.57 (0.08)	1.12 (0.08)	2.5 (0.29)
3	FALSE	0.06 (0.01)	0.13 (0.03)	0.3 (0.04)	0.69 (0.07)	1.49 (0.09)	3.29 (0.16)	7.71 (1.44)
4	FALSE	0.09 (0.01)	0.2 (0.03)	0.54 (0.06)	1.3 (0.11)	3.27 (0.39)	7.13 (0.82)	14.69 (0.96)
5	FALSE	0.15 (0.03)	0.37 (0.06)	1.12 (0.12)	2.84 (0.27)	6.72 (0.5)	17.39 (2.34)	37.64 (9.55)
6	FALSE	0.19 (0.02)	0.46 (0.05)	1.32 (0.12)	3.63 (0.23)	8.75 (0.67)	22.53 (3.34)	76.42 (27.41)
7	FALSE	0.26 (0.02)	0.58 (0.05)	1.74 (0.14)	4.88 (0.36)	12.44 (0.97)	33.34 (5.19)	122.46 (43.26)
2	TRUE	0.05 (0.01)	0.15 (0.06)	0.42 (0.11)	0.51 (0.06)	0.99 (0.05)	2.29 (0.1)	5.29 (0.16)
3	TRUE	0.2 (0.16)	0.3 (0.06)	0.56 (0.08)	1.61 (0.18)	2.38 (0.66)	3.33 (0.04)	8.06 (0.18)
4	TRUE	0.07 (0.01)	0.16 (0.01)	0.37 (0.01)	0.85 (0.02)	1.95 (0.03)	4.64 (0.14)	10.84 (0.27)
5	TRUE	0.09 (0.01)	0.21 (0.01)	0.49 (0.03)	1.13 (0.05)	2.51 (0.08)	5.69 (0.13)	16.54 (0.97)
6	TRUE	0.13 (0.01)	0.31 (0.02)	0.71 (0.04)	1.62 (0.08)	3.66 (0.16)	8.31 (0.21)	19.64 (0.35)
7	TRUE	0.18 (0.03)	0.37 (0.03)	0.87 (0.06)	1.86 (0.07)	4.21 (0.11)	9.74 (0.2)	22.98 (0.33)

TABLE III. Time in seconds to find a minimal experiment for classifying or testing and classifying 100 restricted models consisting of $|X|$ objects and t types, each consisting of three randomly-sampled rankings. Single-core performance on a laptop computer with an Intel i7-1355u processor. Values show the mean (and standard deviation) of computation time in seconds.