Implicit Preferences[*]

Tom Cunningham[†] Jonathan de Quidt[‡]

February 18, 2022

Latest version here

Abstract

We show how simple decisions can, by themselves, reveal two layers of preference. Consider a hiring manager who always chooses a woman over a man with the same qualifications, but always chooses the man if their qualifications differ. Intuitively, these intransitive choices reveal an *explicit* preference for women, but an *implicit* preference for men. More generally, we define an implicit preference for an attribute as one whose influence increases as the attribute is mixed with a superset of other attributes ("dilution"). We show that implicit preferences arise under a diverse set of behavioral foundations: rule-based decision-making, signaling motives, and implicit associations. We prove a representation theorem for the model and show how implicit preferences can be identified from binary choices, or joint evaluation data. We apply the model to two published datasets, finding evidence for implicit risk preferences, implicit selfishness, and implicit discrimination.

[*] We thank for comments, among others, Roland Benabou, Colin Camerer, Ed Glaeser, Sendhil Mullainathan, Pietro Ortoleva, Antonio Rangel, Tomasz Strzalecki, and seminar audiences at USC, Columbia, Caltech, Stanford, Facebook, Harvard, Princeton, Santa Cruz, Gothenburg, Waseda, Berlin, VIBES, and the Stockholm School of Economics. First circulated in 2015 as "Implicit Preferences Inferred from Choice."

[†] Twitter. tom.cunningham@gmail.com

[‡] Institute for International Economic Studies, Stockholm, CAGE, CEPR, CESifo, ThReD. jonathan.dequidt@iies.su.se

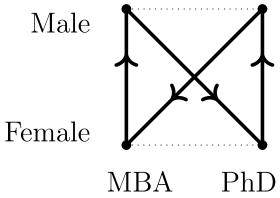
1 Introduction

> *"However we may conceal our passions under the veil, there is always some place where they peep out"* - La Rochefoucauld.

Suppose you observe a hiring manager's choices between pairs of job applicants who differ in gender, and have either an MBA or a PhD. You notice that:

1. They choose the woman when the candidates' qualifications are the same,
2. They choose the man when the candidates' qualifications differ.

Using $A \succ B$ to represent the choice of A from $\{A, B\}$, we can visualize these choices:



The choices are intransitive and therefore inconsistent with standard utility maximization. But they form an intuitive "figure 8" pattern, and seem to reveal two distinct attitudes towards gender, favoring women in the vertical choice sets, and favoring men in the diagonals.

We generalize this observation of choices revealing two distinct preferences. We study preferences over bundles of binary attributes (male/female, Black/white, aisle/window), and we assume that the decision maker has an implicit preference for each attribute (positive, neutral, or negative). We identify implicit preferences from behavior with a "dilution" assumption: the implicit preference for an attribute has more influence when the comparison mixes that attribute with a superset of other attributes. In the example above the diagonal choice sets mix gender with qualification, strengthening the influence of the decision maker's implicit preference for men over women, causing the intransitivity.

The model can also be applied to *evaluation* data, such as willingness to pay, teachers' grading decisions, or judges' sentencing decisions, when the evaluation involves a comparison. Suppose our manager is setting wages for pairs of new hires, one male and the other female. In our model, the manager's implicit preference for men over women will make the *man's* wage sensitive to the *woman's* attributes. For example, we would predict that the man's wage is always lower when the woman has the same qualification as him, compared to when she has a different qualification.

Our formal model presupposes a comparative utility function $u(\boldsymbol{x}, \boldsymbol{z})$, where $\boldsymbol{x}$ is the bundle under consideration (which we will call the "target"), and $\boldsymbol{z}$ is its "comparator." We assume that implicit preferences over attributes are separable and their influence increases when the comparison becomes more "opaque", where opacity is derived from a partial order over the set of all possible comparisons. Our core result is a representation theorem stating that a dataset, consisting of inequalities between comparative utilities, is consistent with a given set of implicit preferences over the attributes if and only if there does not exist a matching between the inequalities such that bundles with implicitly-preferred attributes are ranked higher when opacity is lower, and vice versa.

The theorem makes no assumptions on the partial order applied to opacity. To give the theory content we assume (1) opacity depends only on the set of attributes shared between the target and comparator ($|\boldsymbol{x} - \boldsymbol{z}|$); (2) implicit preferences obey "dilution", meaning the opacity of an attribute is higher in a comparison which mixes that attribute with a superset of other attributes. For evaluation data, we add a third assumption, which permits additional identification when an attribute changes from shared to non-shared.

We then derive a typology of choice and evaluation datasets from which implicit preferences can be immediately read off, or which falsify the model. Each is straightforward to test for empirically. We discuss cases where the data identify the direction of an implicit preference for a given attribute (e.g., an implicit preference for men over women, as in our example), or instead a disjunction between implicit preferences over a set of attributes. Unambiguous identification of a single implicit preference is possible from as few as four binary choices (as in our hiring example), or three joint evaluations.

Our definition of "implicitness" is behavioral, similar to decision-theoretic concepts like "complementarity" or "elasticity" which are defined without reference to the underlying psychology. Why might decisions exhibit implicit preferences? We state three models that exhibit implicit preferences, and satisfy all of our opacity assumptions.

The first foundation (*ceteris paribus*) is a decision maker who is subject to a set of rules that apply in "all else equal" situations, and incurs a utility penalty if they break a rule. When the penalty is infinitely large it represents a hard constraint, and so is a special case of models in which the decision maker chooses from a subset of elements that are maximal by some other set of rankings (e.g. Manzini and Mariotti (2007), Masatlioglu et al. (2012), Cherepanov et al. (2013), Ridout (2021)). Under this model our manager prefers men over women. He can follow his preference when gender is mixed with other attributes, but a rule constrains him from choosing a man over an equally-qualified women, generating a cycle.

The second foundation is a *signaling* model: the decision maker has intrinsic preferences over bundles but also cares about others' perceptions of those preferences. This foundation relates to work on signaling motives, excuse-driven behavior and "moral wiggle room" (e.g. Benabou and Tirole (2003, 2006), Norton et al. (2004), Dana et al. (2007), Andreoni and

Bernheim (2009), Exley (2016)). Choosing a male over a female candidate reveals less about the decision maker's gender preferences as gender is mixed with other attributes, giving them more scope to express their intrinsic preference. Thus, under a signaling interpretation, the manager prefers men but wishes people to believe he has a preference for women.

The third foundation is an *implicit associations* decision maker for whom some knowledge is tacit. This foundation relates to psychological theories of implicit bias and unconscious judgment (e.g. Greenwald and Krieger (2006), Greenwald et al. (1998), Kahneman (2011), Rand et al. (2012)). Under this model the hiring manager holds an unconscious positive association towards men, which gives them a "good feeling" about male candidates. Feelings are typically a good signal of quality, but if they have reason to distrust feelings about gender then this can generate inconsistent behavior. When the candidates differ only in gender the decision maker can diagnose the source of their good feeling and overrule it. As gender is mixed with other attributes, it becomes harder to distinguish the influence of gender from other associations, so their judgment is swayed.

The implicit preference in favor of male candidates revealed by the "figure 8" cycle can thus be interpreted in three ways: (1) a sincere preference for men that is sometimes constrained by rules; (2) a sincere preference for men that is sometimes obscured by signaling motives; (3) an unconscious positive association in favor of men that loses its power when it becomes accessible to conscious awareness.

Finally, we apply the model to two existing datasets. Our first application identifies implicit selfishness and implicit risk attitudes in choice data from Exley (2016), our second identifies implicit racism in evaluation data from DeSante (2013).

We do not know of any prior theory which identifies implicit preferences from multiattribute choice. Existing theories of menu-dependent preferences do not predict the figure 8 pattern.[1] Nevertheless we think that the idea of implicit preferences being revealed by more or less "dilute" decisions taps into a commonsense understanding of decision-making, and most of our formal results correspond to natural intuitions.

A set of related theories are proposed by Manzini and Mariotti (2012) (MM) ("choice by lexicographic semiorder"), Cherepanov et al. (2013) (CFS) ("rationalization"), and Ridout (2021) (R) ("justification"). In these models the decision maker facing a given choice set chooses the element which maximizes their true preference, out of the subset which are

[1] E.g. "salience" (Bordalo et al. (2013)), "relative thinking" (Bushong et al. (2020)), "magnitude effects" (Cunningham (2013)), or "focusing" (Kőszegi and Szeidl (2012)). To the best of our knowledge the only paper besides Cunningham (2014) which identifies a figure 8 pattern in choice is Cubitt et al. (2018) which finds a figure 8 in cross-modal intertemporal tradeoffs. They propose that "the weight put on each attribute of an option is inversely related to how many attributes differ between those options." We discuss in Section 5 how this type of model will not generally exhibit implicit preferences in our sense; Cubitt et al. (2018)'s model does so due to an assumption that money does not have a variable weight.

"justifiable," meaning that the element is undominated relative to at least one of a set of given relations. The models differ on the nature of the relations: MM assume a single semiorder, CFS assume multiple binary relations, and R assumes multiple complete orders. We regard this class of models as complementary to ours. The most important difference is that these models treat outcomes as "atomic" while we treat outcomes as bundles of binary attributes. Models with atomic outcomes are more parsimonious, and those models give unambiguous predictions for choice sets with 3 or more elements, which ours does not.[2]

An advantage of using bundles of attributes is that we can infer implicit preferences from binary choice.[3] Additionally, attributes allow us to make out-of-sample predictions: our hiring manager's gender bias can be predicted to carry over to choice between new candidates with entirely different profiles. Finally, our model can be readily applied to data on joint evaluation, as well as choice. This set of features makes our model well suited to applied work, and in Section 3 we provide an extensive collection of identification tools that can be, and have been, implemented in existing datasets and new experiments. We demonstrate this with our own applications. Our tools have also been adopted by others: Barron et al. (2020) use our approach and find evidence of implicit gender bias.

In psychology, the term "implicit" is usually reserved for cognition, attitudes, judgments, preferences, or knowledge that are "outside conscious attentional focus" (Greenwald and Krieger, 2006), often described as "automatic," "unconscious," "associative." In dual-process theories (e.g., Kahneman (2011)) they are associated with the fast "System 1." In contrast, explicit attitudes are those that are stated or revealed deliberately. Psychologists have developed an array of *non-choice* techniques, most notably, the Implicit Association Test (IAT) (Greenwald et al., 1998), which uses response time to measure implicit associations. IATs have been widely adopted, including within economics (e.g. Glover et al. (2017), Corno et al. (2018), Carlana (2019)), however their ability to predict real-world choices remains controversial (Oswald et al., 2013; Greenwald et al., 2015). In contrast our model defines implicit preferences directly from and with reference to real decisions.

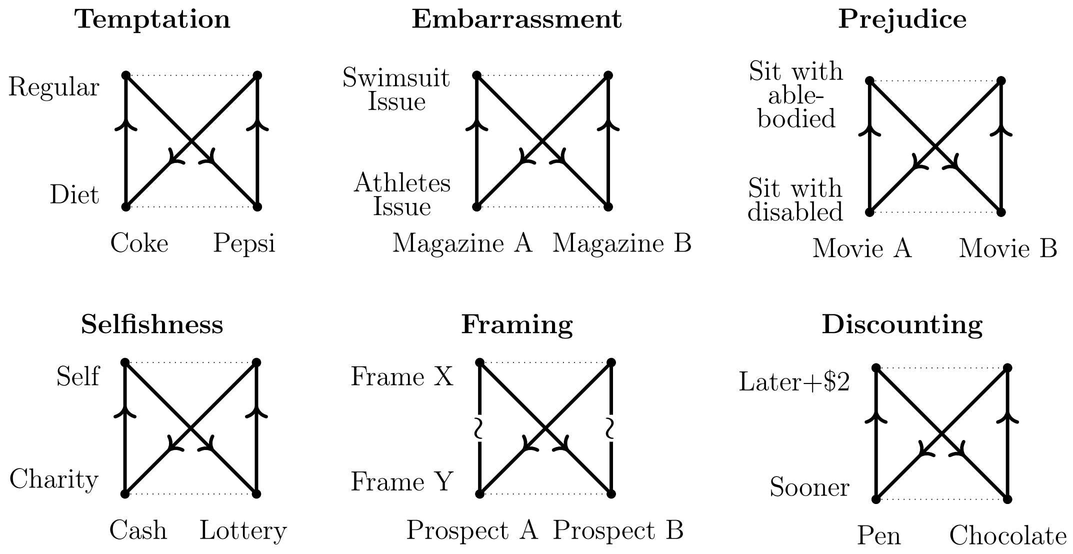
A few empirical studies rely on related intuition to our formal model: Snyder et al. (1979) on discrimination against the disabled, Exley (2016) on excuses for selfish behavior, Bohnet et al. (2016) on gender discrimination, Cubitt et al. (2018) on time discounting. Each paper uses identification approaches tailored to their setting.[4] Our framework is designed to be

[2] Extending the notion of an implicit preference to 3-element comparisons requires additional assumptions. This can be seen because our three foundations, while they coincide on 2-element sets, give different prediction for 3-element sets.

[3] With atomic elements binary choice will generally be uninformative: a cycle of the form $a \succ b \succ c \succ a$ implies only that there must exist some constraint on choice.

[4] Snyder et al. (1979) is also vulnerable to alternative explanations, see Section 6.

generally and broadly applicable.



Temptation. The decision maker chooses between diet and full-sugar sodas. They explicitly prefer diet soda, but reveal an implicit preference for the less healthy option.

Embarrassment. The decision maker chooses between magazines, which may have a swimsuit issue or a special issue covering famous athletes (Chance and Norton, 2009). They reveal an explicit preference for the athletes issue but an implicit preference for the swimsuit issue.

Prejudice. The decision maker chooses between movies, which will be watched with an able-bodied or a disabled person (Snyder et al., 1979). They explicitly prefer to sit with the disabled person, but reveal an implicit preference for sitting with the able-bodied person.

Selfishness. The decision maker chooses between a lottery and a safe amount, where the beneficiary is themself or charity (Exley, 2016). They explicitly prefer to give to charity, but reveal an implicit preference for self, i.e. they are implicitly selfish.

Framing. The decision maker chooses between prospects (A and B) framed in different ways (X and Y). They are indifferent between differently-framed versions of the same prospect, but strictly prefer frame X when the prospects differ. This reveals an implicit preference for frame X, but no explicit preference.

Discounting The decision maker chooses between a pen or a box of chocolates, either now, or with a financially-compensated delay (Cubitt et al., 2018). They reveal an explicit preference for sooner rewards, but an implicit preference for later; i.e. they are implicitly patient.

Figure 1: Figure 8 intransitivities applied to various domains.

Our introductory example shows how we can identify implicit discrimination, a topic of great recent interest.[5] There are many other possible applications, in principle we can detect implicit preferences over any binary attribute, and there are many contexts in which we might expect them. Figure 1 shows a variety of figure 8 cycles in different domains to illustrate implicit preferences that we believe are plausibly detectible.

[5] Bertrand et al. (2005) and Bertrand and Duflo (2017) discuss the economic importance of implicit discrimination, and the difficulty of measuring it. They mention that implicit discrimination will be more pronounced in more "ambiguous" situations: our paper can be seen as formalizing this notion.

2 Model

We will work with a utility function with two arguments, $u(\boldsymbol{x}, \boldsymbol{z})$. Both are bundles of n binary attributes (male/female, PhD/MBA, aisle/window), and we refer to the first as the "target" and the second as the "comparator".

We will assume that the comparator $\boldsymbol{z}$ affects the utility of $\boldsymbol{x}$ through a comparison $\delta(\boldsymbol{x}, \boldsymbol{z}) \in \Delta$, which in turn affects attribute-specific utility weights $\theta_i(\delta(\boldsymbol{x}, \boldsymbol{z}))$ for attributes $i \in \{1 \dots n\}$. We can then express the consistency of a set of implicit preferences with a dataset as the existence of a solution to a set of linear inequalities. Then, applying a theorem of the alternative, we show that is equivalent to the existence of a matching between elements of the dataset.

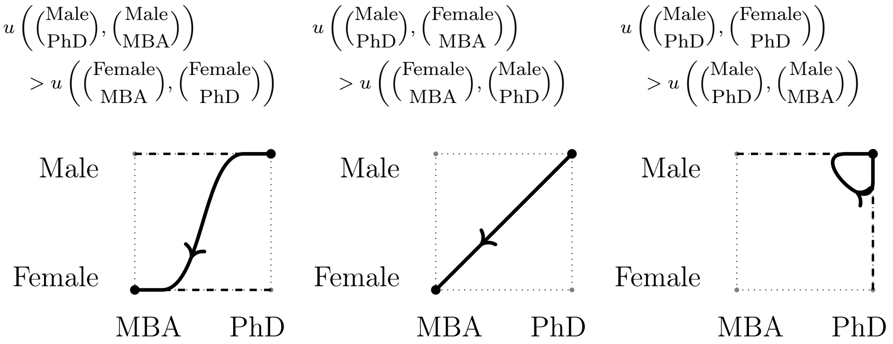
Our representation theorem holds for any definition of $\delta(\boldsymbol{x}, \boldsymbol{z})$. For applications we will assume that a comparison is the absolute value of the difference between bundles, $\delta(\boldsymbol{x}, \boldsymbol{z}) = |\boldsymbol{x} - \boldsymbol{z}|$. Intuitively this means that the utility weights depend just on the set of attributes shared and non-shared between the target and comparator. We then introduce a "dilution" assumption on opacity, which is the key principle we use to identify implicit preferences, and is common to the three foundations in section 4. The dilution assumption says that an implicitly-preferred attribute will get higher weight in comparisons which mix that attribute with a superset of other attributes.

All vectors will be column vectors, indicated with a bold font, and $\boldsymbol{x}^T$ will refer to the transpose of $\boldsymbol{x}$. The absolute value of a vector will be written as $|\boldsymbol{x}|^T = \Big(|x_1| \quad \dots \quad |x_n|\Big)$. Inequalities between vectors will be defined as:

$$\begin{aligned}
\boldsymbol{x} \geq \boldsymbol{z} &\iff \quad x_i \geq z_i \text{ for } i = 1, \dots, n. \\
\boldsymbol{x} > \boldsymbol{z} &\iff \quad x_i \geq z_i \text{ for } i = 1, \dots, n, \text{ and } \boldsymbol{x} \neq \boldsymbol{z}. \\
\boldsymbol{x} \gg \boldsymbol{z} &\iff \quad x_i > z_i \text{ for } i = 1, \dots, n.
\end{aligned}$$

The primitives of utility are **bundles** of n binary attributes: $\boldsymbol{x} \in \mathcal{X} = \{-1, 1\}^n$. A **comparative utility function** is a function $u : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and data will consist of a set of inequalities between comparative utilities, with the form $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}', \boldsymbol{z}')$. This structure allows us to represent data from both binary choice and pairwise evaluation, discussed in section 2.1.

We will frequently use diagrams to visualize sets of inequalities. An arrow from $\boldsymbol{x}$ to $\boldsymbol{x}'$ shows the sign of the inequality, with the entry and exit angles pointing towards each bundle's comparator, $\boldsymbol{z}$ and $\boldsymbol{z}'$. We use dashed lines to show the comparisons: they run from $\boldsymbol{x}$ to $\boldsymbol{z}$ and from $\boldsymbol{x}'$ to $\boldsymbol{z}'$. Figure 2 gives three examples.


$u\left(\binom{\text{Male}}{\text{PhD}},\binom{\text{Male}}{\text{MBA}}\right) > u\left(\binom{\text{Female}}{\text{MBA}},\binom{\text{Female}}{\text{PhD}}\right)$
$u\left(\binom{\text{Male}}{\text{PhD}},\binom{\text{Female}}{\text{MBA}}\right) > u\left(\binom{\text{Female}}{\text{MBA}},\binom{\text{Male}}{\text{PhD}}\right)$
$u\left(\binom{\text{Male}}{\text{PhD}},\binom{\text{Female}}{\text{PhD}}\right) > u\left(\binom{\text{Male}}{\text{PhD}},\binom{\text{Male}}{\text{MBA}}\right)$
Male
Female
MBA
PhD
Male
Female
MBA
PhD
Male
Female
MBA
PhD

In words, the first inequality shows that the utility of a Male PhD compared to a Male MBA is higher than the utility of a Female MBA compared to a Female PhD.

Figure 2: Examples of graphical representations of inequalities.

A **dataset** D is a collection of m 4-tuples, $(\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j)_{j=1}^m$, with $x^j, z^j, x'^j, z'^j \in \mathcal{X}$, and some $\bar{m} \in \mathbb{N}$, $1 \leq \bar{m} \leq m$. We say a comparative utility function u **rationalizes** a dataset D if all m elements of the dataset can be interpreted as inequalities that are satisfied by u, i.e. for every $j = 1, \ldots m$,

$$\begin{aligned} u(\boldsymbol{x}^j, \boldsymbol{z}^j) &> u(\boldsymbol{x}'^j, \boldsymbol{z}'^j) \quad , 1 \leq j \leq \bar{m} \qquad \text{(strict inequalities)} \\ u(\boldsymbol{x}^j, \boldsymbol{z}^j) &\geq u(\boldsymbol{x}'^j, \boldsymbol{z}'^j) \quad , \bar{m} < j \leq m. \quad \text{(weak inequalities)} \end{aligned}$$

We next introduce assumptions on the comparative utility function.

Assumption 1. *The utility of bundle $\boldsymbol{x} \in \mathcal{X}$ with comparator $\boldsymbol{z} \in \mathcal{X}$, is:*

$$u(\boldsymbol{x}, \boldsymbol{z}) = \overbrace{v(\boldsymbol{x})}^{\textit{explicit value}} + \overbrace{\sum_{i=1}^{n} x_i \cdot \underbrace{\kappa_i}_{\substack{\textit{implicit}\\ \textit{preference}\\ \textit{for } i}} \cdot \underbrace{\theta_i(\delta(\boldsymbol{x}, \boldsymbol{z}))}_{\substack{\textit{opacity of}\\ \textit{comparison}\\ \textit{for } i}}}^{\textit{implicit value}}, \tag{1}$$

with $v : \mathcal{X} \to \mathbb{R}$, $\kappa_i \in \{-1, 0, 1\}$, $\delta : \mathcal{X} \times \mathcal{X} \to \Delta$, $\theta_i : \Delta \to \mathbb{R}$.[6]

The implicit value of attribute x_i depends on two factors: (1) the implicit preference, which can be positive ($\kappa_i = 1$), negative ($\kappa_i = -1$) or neutral ($\kappa_i = 0$), and (2) the "opacity" of the comparison for i, $\theta_i(\delta(\boldsymbol{x}, \boldsymbol{z}))$.

[6] Since the dataset consists only of ordinal utilities, we could equivalently a) wrap the utility function in a strictly increasing function (i.e. linearity is not critical), and b) normalize opacity to, e.g. take only non-negative values.

The opacity function $\theta(\cdot)$ maps comparisons to scalar weights, so we can say that a comparison δ has a higher or lower opacity with respect to attribute i, causing a higher or lower weight on i's implicit preference κ_i.

The economic interpretation of an implicit preference κ_i is linked to the polarity of its attribute x_i. Suppose gender is defined as $\text{Male} = 1, \text{Female} = -1$. A positive implicit preference ($\kappa_i = 1$) means that the value of bundles with $x_i = 1$ is increasing in opacity θ_i, while the value of bundles with $x_i = -1$ is decreasing in θ_i. Hence in this example a positive implicit preference on the gender attribute translates into an implicit preference for men over women, whose influence grows as θ_i increases.

Variation in opacity derives from a partial order $\sqsupseteq_i$ over the set of comparisons Δ. Given a pair of comparisons $\delta, \delta' \in \Delta$ we describe $\delta \sqsupseteq_i \delta'$ as δ **opacity dominates** δ' **on attribute** i. We assume that $\theta_i(\cdot)$ obeys this partial order:

Assumption 2 (Opacity Dominance)**.** *For any $\delta, \delta' \in \Delta$,*

$$\delta \sqsupseteq_i \delta' \implies \theta_i(\delta) \geq \theta_i(\delta').$$

Our theorem will give conditions under which a dataset D will be consistent with a set of implicit preferences $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$, given the partial order $\sqsupseteq_i$. We postpone making further assumptions on $\sqsupseteq_i$ until subsection 2.2.

Before coming to the theorem we define some features of a dataset that make the rationalizability condition easier to state.

A **cyclical selection** is a weighted subset of a dataset that satisfies a condition that will be shown to be necessary and sufficient for the data to be inconsistent with a menu-independent utility function $u(\boldsymbol{x}, \boldsymbol{z}) = v(\boldsymbol{x})$. Intuitively it picks out a subset of inequalities that consists entirely of strict cycles.

Definition 1 (Cyclical Selection)**.** *Given a dataset $D = \{\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j\}_{j=1}^m$ a* ***cyclical selection*** *is vector of non-negative integer weights $\boldsymbol{s} \in \mathbb{N}^m$ such that each bundle appears equally often on the left-hand and right-hand sides: i.e., for every $\boldsymbol{x} \in \mathcal{X}$,*

$$\underbrace{\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\}}_{\text{appearances of } \boldsymbol{x} \text{ on LHS}} = \underbrace{\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}}_{\text{appearances of } \boldsymbol{x} \text{ on RHS}},$$

with $s_j > 0$ for at least one $j \in 0, \ldots, \bar{m}$ (i.e., at least one strict inequality is included).

Figure 3 shows three cyclical selections (assuming $s_j = 1$ for each inequality).[7]

[7] When every inequality in a cyclical selection has the form $u(\boldsymbol{a}, \boldsymbol{b}) \geq u(\boldsymbol{b}, \boldsymbol{a})$ then it will consist of a

$$u\left(\binom{1}{1},\binom{-1}{1}\right) > u\left(\binom{1}{1},\binom{1}{-1}\right)$$

$$u\left(\binom{1}{1},\binom{-1}{-1}\right) > u\left(\binom{-1}{-1},\binom{1}{1}\right)$$
$$u\left(\binom{-1}{-1},\binom{-1}{1}\right) > u\left(\binom{-1}{1},\binom{-1}{-1}\right)$$
$$u\left(\binom{-1}{1},\binom{1}{1}\right) > u\left(\binom{1}{1},\binom{-1}{1}\right)$$

$$u\left(\begin{pmatrix}1\\1\\1\end{pmatrix},\begin{pmatrix}-1\\-1\\-1\end{pmatrix}\right) > u\left(\begin{pmatrix}-1\\-1\\-1\end{pmatrix},\begin{pmatrix}1\\1\\1\end{pmatrix}\right)$$
$$u\left(\begin{pmatrix}-1\\-1\\-1\end{pmatrix},\begin{pmatrix}-1\\1\\-1\end{pmatrix}\right) > u\left(\begin{pmatrix}-1\\1\\-1\end{pmatrix},\begin{pmatrix}-1\\-1\\-1\end{pmatrix}\right)$$
$$u\left(\begin{pmatrix}-1\\1\\-1\end{pmatrix},\begin{pmatrix}1\\1\\1\end{pmatrix}\right) > u\left(\begin{pmatrix}1\\1\\1\end{pmatrix},\begin{pmatrix}-1\\1\\-1\end{pmatrix}\right)$$

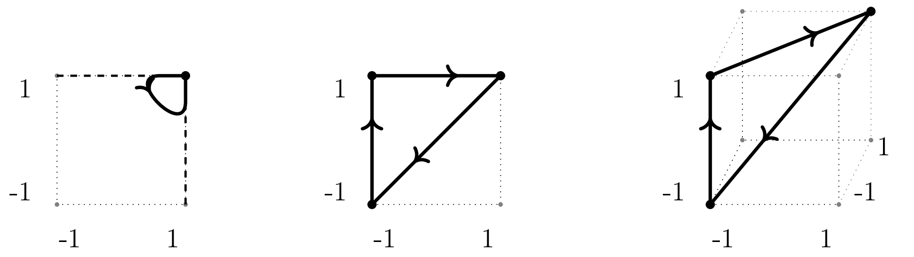


Figure 3: Examples of cyclical selections with two or three attributes

Finally we define "wins," "losses," and "score". Given an inequality $u(\boldsymbol{x},\boldsymbol{z}) > u(\boldsymbol{x}',\boldsymbol{z}')$ we count it as a "win" for attribute i if the preferred bundle, $\boldsymbol{x}$, has a positive realization of that attribute ($x_i = 1$), or the dispreferred bundle, $\boldsymbol{x}'$, has a negative realization ($x'_i = -1$). We define a "loss" as the reverse. Thus, if gender is encoded as Male = 1, it is a win if a man appears on the LHS or a woman on the RHS, and a loss if a woman appears on the LHS or a man on the RHS. For a given comparison δ, the "score" will be the net number of wins.

Definition 2 (Score)**.** *Given a dataset $D = \{x^j, z^j, x'^j, z'^j\}_{j=1}^m$ and a selection $\boldsymbol{s} \in \mathbb{N}^m$ the* ***score*** *vector, $\boldsymbol{c} \in \mathbb{Z}^{n|\Delta|}$ represents for each $i \in \{1,\ldots,n\}$ and $\delta \in \Delta$, the net number of times that the attribute wins in $\boldsymbol{s}$:*

$$c_{i,\delta} = \underbrace{\sum_{\{j:\delta(\boldsymbol{x}^j,\boldsymbol{z}^j)=\delta\}} s_j x_i^j}_{\textit{bundles on LHS of inequality}} - \underbrace{\sum_{\{j:\delta(\boldsymbol{x}'^j,\boldsymbol{z}'^j)=\delta\}} s_j x_i'^j}_{\textit{bundles on RHS of inequality}} .$$

In a cyclical selection each bundle appears equally often on the LHS and RHS, so the score must sum to zero across comparisons for a given attribute i: $\sum_{\delta\in\Delta} c_{i,\delta} = 0$.

Keeping track of wins and losses allows us to test whether implicit preferences can rationalize a dataset. A positive implicit preference for some attribute cannot rationalize a cycle if opacity is always weakly higher in losses than in wins (and a negative implicit preference cannot rationalize it if opacity is weakly higher for wins than losses). We formalize this condition with the existence of a matrix M_i which matches scores, effectively matching wins

collection of intransitive cycles between the target bundles (the first arguments in the utility function). This follows from Euler's theorem (Jungnickel (2005)): if a connected multigraph has equal in-degree and out-degree then the edges can be partitioned into cycles.

to losses, in which each match satisfies the relation $\sqsupseteq_i$, meaning we can rank the opacity of the wins and losses.

Definition 3 (Opacity Dominance)**.** *Given a vector of scores for attribute i, $\boldsymbol{c}_i \in \mathbb{Z}^{\Delta}$ we say **wins opacity dominate losses** if there exists a matrix $M_i \in \mathbb{N}^{\Delta\times\Delta}$ with:*

$$\begin{aligned}
&\forall \bar\delta, \bar\delta' \in \Delta, \quad (M_{i,\bar\delta,\bar\delta'} > 0) \implies (\bar\delta \sqsupseteq_i \bar\delta') && \textit{(matches obey dominance)} \\
&\forall \delta \in \Delta, \qquad c_{i,\delta} = \underbrace{\sum_{\bar\delta'\in\Delta} M_{i,\delta,\bar\delta'}}_{\substack{\textit{outging matches} \\ (\delta \textit{ dominating})}} - \underbrace{\sum_{\bar\delta\in\Delta} M_{i,\bar\delta,\delta}}_{\substack{\textit{incoming matches} \\ (\delta \textit{ dominated})}} && \textit{(all scores are accounted for)}
\end{aligned}$$

The second condition checks that all scores are matched: each positive score, $c_{i,\delta} > 0$, where wins exceed losses, will have a net outflow equal to $c_{i,\delta}$, and each negative score, $c_{i,\delta} < 0$ will have a net inflow equal to $|c_{i,\delta}|$.

We likewise say that **losses opacity dominate wins** if there is an M that satisfies the same conditions but the last line in the definition sums to $-c_{i,\delta}$ instead of $c_{i,\delta}$.

The second condition shows that matrix M_i can be thought of as a *matching* between scores in $\boldsymbol{c}_i$: the net inflow for each element δ (the incoming minus outgoing matches) is equal to the score for that element, $c_{i,\delta}$.

We are now ready to state our representation theorem. It tells us that a candidate vector of implicit preferences $\boldsymbol{\kappa}$ can rationalize the data if and only if there does not exist any matching between wins and losses that contradicts the proposed signs of every κ_i in the vector. Thus the set of implicit preferences that can rationalize the data is the set of $\boldsymbol{\kappa}$ vectors that fulfill this condition.

Theorem 1 (Representation)**.** *A dataset D can be rationalized with implicit preferences $\boldsymbol{\kappa}$ if and only if it does not contain a cyclical selection $\boldsymbol{s}$ such that, for every attribute with a positive implicit preference ($\kappa_i = 1$), the losses opacity dominate the wins, and for every attribute with a negative implicit preference ($\kappa_i = -1$), the wins opacity dominate the losses.*

2.1 Choice and Evaluation as Datasets

We have defined a dataset as a set of inequalities between comparative utilities. We now show how data from choice and evaluation can be represented in a dataset.

For binary choice we treat each bundle as the other bundle's comparator: if someone expresses a strict preference for $\boldsymbol{x}$ over $\boldsymbol{z}$ we treat that as an inequality, $u(\boldsymbol{x},\boldsymbol{z}) > u(\boldsymbol{z},\boldsymbol{x})$. If someone expresses indifference we can represent that as a pair of weak inequalities, $u(\boldsymbol{x},\boldsymbol{z}) \geq u(\boldsymbol{z},\boldsymbol{x})$ and $u(\boldsymbol{z},\boldsymbol{x}) \geq u(\boldsymbol{x},\boldsymbol{z})$. Thus a choice cycle $\boldsymbol{x}^a \succ \boldsymbol{x}^b \succ \boldsymbol{x}^c \succeq \boldsymbol{x}^a$ will correspond to a dataset with $m = 4, \bar{m} = 3$, written:

$$\begin{aligned} u(\boldsymbol{x}^a, \boldsymbol{x}^b) &> u(\boldsymbol{x}^b, \boldsymbol{x}^a) \\ u(\boldsymbol{x}^b, \boldsymbol{x}^c) &> u(\boldsymbol{x}^c, \boldsymbol{x}^b) \\ u(\boldsymbol{x}^c, \boldsymbol{x}^a) &\geq u(\boldsymbol{x}^a, \boldsymbol{x}^c) \\ u(\boldsymbol{x}^a, \boldsymbol{x}^c) &\geq u(\boldsymbol{x}^c, \boldsymbol{x}^a). \end{aligned}$$

Theorem 1 is also applicable to data on continuous *evaluations* of bundles when each evaluated bundle has a comparator. We assume that evaluations are a strictly increasing function of utility: $y(\boldsymbol{x}, \boldsymbol{z}) = f(u(\boldsymbol{x}, \boldsymbol{z})), f' > 0$. To apply our theorem we construct a set of inequalities sufficient to represent the ordinal relationship of the evaluations: we first rank each evaluation (breaking ties arbitrarily), and enter an inequality into the dataset for each pair of consecutive evaluations. When two evaluations are equal we use two weak inequalities going in opposite directions. For example suppose we observe the three willingness-to-pay judgments for bundles with comparators: $y(\boldsymbol{x}^a, \boldsymbol{x}^b) = \$310$, $y(\boldsymbol{x}^a, \boldsymbol{x}^c) = \$200$, $y(\boldsymbol{x}^b, \boldsymbol{x}^a) = \$200$. Then we would construct a dataset with three inequalities:

$$\begin{aligned} u(\boldsymbol{x}^a, \boldsymbol{x}^b) &> u(\boldsymbol{x}^a, \boldsymbol{x}^c) \\ u(\boldsymbol{x}^a, \boldsymbol{x}^c) &\geq u(\boldsymbol{x}^b, \boldsymbol{x}^a) \\ u(\boldsymbol{x}^b, \boldsymbol{x}^a) &\geq u(\boldsymbol{x}^a, \boldsymbol{x}^c). \end{aligned}$$

To apply the theorem, one must search over *all* cyclical selections that can be constructed from the dataset to verify that the matching condition holds. For simple datasets such as a single choice cycle this is straightforward and can often be done by simple visual inspection. We provide numerous canonical examples of datasets, along with their implied implicit preferences, in Section 3. For larger datasets the matching procedure can be laborious. However, the proof of Theorem 1 uses an equivalency result between the matching procedure and a matrix representation of the dataset, that can then be solved numerically.

2.2 Assumptions on Comparisons and Opacity

We now add assumptions on comparisons ($\delta(\cdot)$) and on opacity ($\sqsupseteq_i$) to tailor our representation theorem to identify implicit preferences.

Assumption 3 (Equivalence)**.** *We define the set of comparisons $\Delta = \{0, 2\}^n$, and the comparison function, $\delta : \mathcal{X} \times \mathcal{X} \to \Delta$ as,*

$$\delta(\boldsymbol{x}, \boldsymbol{z}) = |\boldsymbol{x} - \boldsymbol{z}|.$$

From this point we will treat the comparison $\boldsymbol{\delta}$ as a vector (and so print it in bold) with $\delta_i = |x_i - z_i|$. We will use the term **status** to refer to whether an attribute is shared ($\delta_i = 0$) or non-shared ($\delta_i = 2$) in a comparison.[8]

Our primary assumption on $\theta_i(\cdot)$ will be that an attribute i becomes more opaque as additional other attributes share status with i. We call the assumption "dilution."

Suppose $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on gender. Dilution implies that an implicit preference for one gender over the other will be weakest when $\boldsymbol{x}$ and $\boldsymbol{z}$ differ *only* on gender, and become progressively stronger as $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on more other attributes in addition to gender. This is the key property that allows us to identify implicit preferences from choice and evaluation data, and is shared by our three foundations.

Assumption 4 (Dilution)**.** *For any $i \in \{1, \ldots, n\}, \boldsymbol{\delta}, \boldsymbol{\delta}' \in \Delta$,*

$$\underbrace{(\delta_i = \delta'_i)}_{\substack{\textit{attribute } i \textit{ has} \\ \textit{same status in } \delta \textit{ \& } \delta'}} \quad \wedge \quad \underbrace{\{j : \delta_j = \delta_i\} \supseteq \{j : \delta'_j = \delta'_i\}}_{\substack{\textit{more attributes share status with } i \\ \textit{in } \delta \textit{ than in } \delta'}} \quad \Longrightarrow \quad \underbrace{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'}_{\substack{\textit{attribute } i \textit{ is more} \\ \textit{opaque under } \delta \textit{ than } \delta'}}$$

Assumptions 1–4 constitute our workhorse model, "Separable Implicit Preferences."

Definition 4 (Separable Implicit Preferences)**.** *A* Separable Implicit Preferences *decision maker satisfies Assumptions 1, 2, 3, and 4.*

Assumptions 1–4 are sufficient for all of our analysis of choice data and for a number of identification results in evaluation data. But for some evaluation results we need to be able to rank opacities for attributes that change status, from shared to non-shared or vice versa, and Dilution has nothing to say in these cases.[9] Our final assumption allows us to do this. It assumes that one attribute, $k \in \{1, \ldots, n\}$, is "special" in the sense that opacity is greater for attributes that have the same status as k. If k is shared, then opacity is greater for shared attributes, if k is non-shared, opacity is greater for non-shared attributes.

Assumption 5 (Dominance of attribute k)**.** *For any $i \in \{1, \ldots, n\} \setminus k, \boldsymbol{\delta}, \boldsymbol{\delta}' \in \Delta$,*

$$\underbrace{\delta_i = \delta_k \wedge \delta'_i \neq \delta'_k}_{\substack{\textit{same status as } k \textit{ in } \delta \\ \textit{diff status from } k \textit{ in } \delta'}} \implies \underbrace{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'}_{\substack{\textit{more opaque} \\ \textit{under } \delta \textit{ than } \delta'}}$$

[8] If we define comparison in a more granular way then opacity can depend on the direction of the comparison ($\theta(\boldsymbol{x} - \boldsymbol{z})$), or the identity of the target bundle ($\theta(\boldsymbol{x}, \boldsymbol{z})$). Those broader definitions can accommodate richer models, but the definition adopted here is sufficient for the three foundational models we describe in section 4.

[9] Dilution is sufficient for analysis of choice data because implicit preferences on shared attributes have no influence on choice.

We think it is most natural to think of k as a shared attribute, capturing what is "held constant" in the decision.[10] In section 4 we provide conditions under which each foundation implies Assumption 5. For the foundations based on signal extraction (signaling and implicit associations) the intuition is that there is high uncertainty about the value of attribute k, such that little can be learned about attributes share status with k.

3 Canonical Examples

We now work through a set of important classes of dataset, from both choice and evaluation, picked to encompass those that are most useful for applications. We state each as a corollary of Theorem 1. The proofs are mechanical so we consign them to web appendix B.1. Definitions are stated in terms of strict inequalities, however all results will go through provided at least one inequality in each cycle is strict.

For choice, we introduce the *right triangle*, the shortest choice cycle (three choices) that provides unambiguous restrictions on $\boldsymbol{\kappa}$. It provides a disjunction over the implicit preference on all attributes that vary in the cycle. Second, we discuss the *figure 8*, which is the shortest choice cycle (four choices) that can uniquely identify a single implicit preference. Third, we show how pairs of *parallel right triangles* (five or six choices) can provide further restrictions relative to the single triangle, including the possibility of unique identification.

Turning to evaluation we introduce the *convex scissor*, the smallest cyclical selection (two joint evaluations) that can provide unambiguous restrictions on *some elements of* $\boldsymbol{\kappa}$. It implies the existence of at least one implicit preference, but only restricts the sign of a subset. Next, we show how pairs of *parallel convex scissors* (three or four joint evaluations) can refine identification relative to the single scissor, including the possibility of unique identification. We also show how Assumption 5 (Dominance of attribute k) further refines identification.

Finally, we present two examples that reveal the presence of an implicit preference but no further restrictions, and three that falsify the theory.

For each corollary we provide three examples with two or three attributes. We state each example's implications for $\boldsymbol{\kappa}$ and in natural language. E.g., $\kappa_1 > 0$ means we learn κ_1 is positive, $\kappa_1 \neq 0$ means we learn there is an implicit preference for attribute 1 but not its

[10] Consider a judge assigning sentences to two defendants simultaneously. If both defendants are Black, they receive long sentences. If both are white, they receive short sentences. But if one is Black and the other is white, they receive intermediate sentences. Under the assumption that opacity is higher for shared attributes, this reveals an implicit preference for white and against Black defendants. Intuitively, the long sentence assigned to two Black defendants could be explained by their race, but also by prevailing sentencing rules, the leniency of the judge, the time of day, and so on. These shared attributes cannot explain why sentences differ between otherwise similar Black and white defendants. Hence, opacity about race is greater when race is shared than when it is non-shared. Our DeSante (2013) application has this structure.

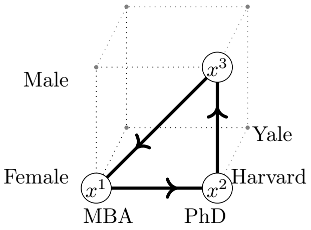
sign, and so on. In natural language, we always state preferences relative to the +1 pole of the attribute. +Male means an implicit preference favoring men (relative to women), −Male means an implicit preference favoring women (i.e., against men), and ±Male means we learn there is an implicit gender preference not its sign.

Choice examples. We begin with choice data. We assume throughout that the decision maker satisfies Separable Implicit Preferences (Assumptions 1–4).

Definition 5 (Right triangle)**.** *A right triangle is a choice cycle over three distinct bundles, ordered* $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$*, with corresponding differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \boldsymbol{\delta}^3$*, in which the differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2$ *are orthogonal, i.e.* $\boldsymbol{\delta}^1$ *and* $\boldsymbol{\delta}^2$ *differ on distinct sets of attributes.*

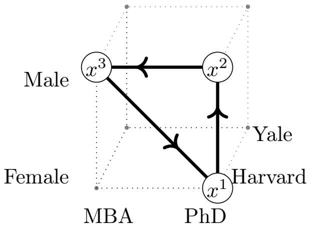
Corollary 1. *A right triangle implies at least one nonzero implicit preference favoring* $\boldsymbol{x}^3$*'s realization of an attribute on which it differs from* $\boldsymbol{x}^1$*:*

$$\bigvee_{i:x_i^3 \neq x_i^1} (x_i^3 \kappa_i = 1).$$



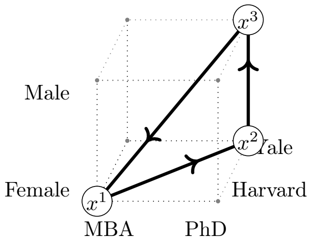
$(\kappa_1 > 0) \vee (\kappa_2 > 0)$

+PhD ∨ +Male.



$(\kappa_1 < 0) \vee (\kappa_2 > 0)$

−PhD ∨ +Male.



$(\kappa_1 > 0) \vee (\kappa_2 > 0) \vee (\kappa_3 > 0)$

+PhD ∨ +Male ∨ +Yale.

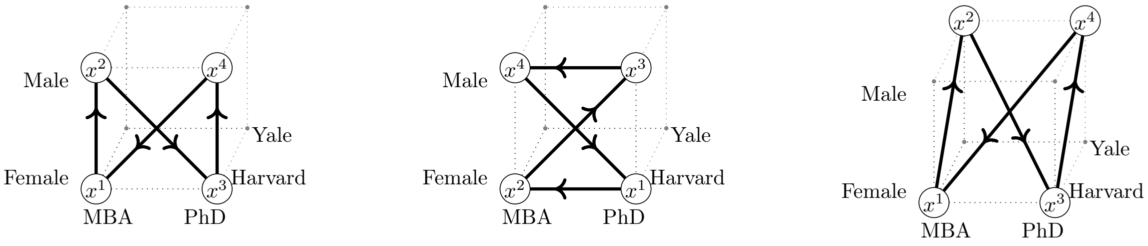
A single right triangle cannot uniquely identify an implicit preference without further restrictions on $\boldsymbol{\kappa}$, because x^3 and x^1 must differ on at least two attributes.

For intuition note that the first right triangle can be thought of as containing two reversals of preference: (1) a female PhD is chosen over a male PhD, but the gender preference is reversed when the man has an MBA (which dilutes the gender attribute); (2) a female MBA is chosen over a female PhD, but the qualification preference is reversed when the PhD holder is male (which dilutes the qualification attribute). The cycle thus implies the presence of at least one implicit preference but we cannot distinguish between one favoring men, one favoring PhDs, or both.

Definition 6 (Figure 8)**.** *A figure 8 is a choice cycle over four distinct bundles, ordered* $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^4 \succ \boldsymbol{x}^1$*, with corresponding differences* $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \boldsymbol{\delta}^3, \boldsymbol{\delta}^4$*. It must satisfy two conditions: (1) there are only two sets of differences* $\boldsymbol{\delta}^1 = \boldsymbol{\delta}^3$ *and* $\boldsymbol{\delta}^2 = \boldsymbol{\delta}^4$*; and (2) the even-numbered comparisons differ on a superset of attributes:* $\boldsymbol{\delta}^2 > \boldsymbol{\delta}^1$*.*

Corollary 2. *A figure 8 implies at least one nonzero implicit preference, favoring $\boldsymbol{x}^4$'s realization of an attribute on which it differs from $\boldsymbol{x}^3$:*

$$\bigvee_{i:x_i^3 \neq x_i^4} (x_i^4 \kappa_i = 1).$$



$(\kappa_2 > 0)$
+Male.

$(\kappa_1 < 0)$
−PhD.

$(\kappa_2 > 0) \vee (\kappa_3 > 0)$
+Male ∨ +Yale.

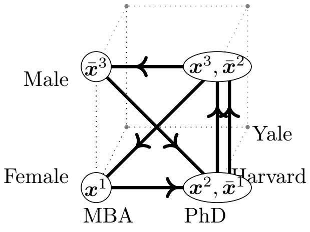
Since $\boldsymbol{x}^3$ and $\boldsymbol{x}^4$ may differ on only one attribute, a figure 8 cycle can uniquely identify an implicit preference.

The figure 8 can be thought of as containing two preference reversals. In the leading example, a female candidate is chosen over a man with the same qualification, but is rejected whenever the qualifications differ (which dilutes the gender attribute). One reversal favors male MBAs, the other favors male PhDs. Under our assumptions, an implicit preference on the qualification dimension cannot generate both, so there must be an implicit preference favoring men.

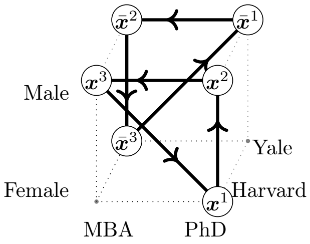
Definition 7 (Parallel right triangles)**.** *A pair of parallel right triangles is a selection consisting of two right triangles $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$ and $\bar{\boldsymbol{x}}^1 \succ \bar{\boldsymbol{x}}^2 \succ \bar{\boldsymbol{x}}^3 \succ \bar{\boldsymbol{x}}^1$, which satisfy two conditions: (1) identical* signed *differences on $\{\boldsymbol{x}^2, \boldsymbol{x}^3\}$ and $\{\bar{\boldsymbol{x}}^1, \bar{\boldsymbol{x}}^2\}$ (that is, $\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2$, implying $\boldsymbol{\delta}^2 = \bar{\boldsymbol{\delta}}^1$); and (2) opposing* signed *differences on $\{\boldsymbol{x}^1, \boldsymbol{x}^2\}$ and $\{\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3\}$ (that is, $\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3)$, implying $\boldsymbol{\delta}^1 = \bar{\boldsymbol{\delta}}^2$).*

Corollary 3. *A pair of parallel right triangles implies at least one implicit preference, favoring $\boldsymbol{x}^3$'s realization of an attribute on which it differs from $\boldsymbol{x}^2$:*

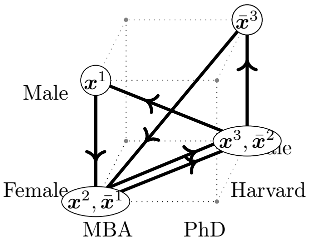
$$\bigvee_{i:x_i^3 \neq x_i^2} (x_i^3 \kappa_i = 1).$$



$(\kappa_2 > 0)$
+Male.



$(\kappa_1 < 0)$
−PhD.



$(\kappa_1 > 0) \vee (\kappa_3 > 0)$
+PhD ∨ +Yale.

As $\boldsymbol{x}^3$ and $\boldsymbol{x}^2$ may differ on only one attribute, a pair of parallel right triangles can uniquely identify an implicit preference. They achieve this by ruling out part of each triangle's individual disjunctions. In particular, they eliminate all attributes that vary in $\boldsymbol{\delta}^1/\bar{\boldsymbol{\delta}}^2$ (the attributes over which the triangles have opposing preferences), leaving only the attributes that vary in $\boldsymbol{\delta}^2/\bar{\boldsymbol{\delta}}^1$ (where the triangles agree).

Evaluation examples. We now turn to evaluation data. In each case we first state the result under the assumption of Separable Implicit Preferences (Assumptions 1–4). Then we show how adding Assumption 5 can refine identification.

Definition 8 (Convex scissor)**.** *A convex scissor is a pair of evaluations of a single bundle* $\boldsymbol{x}$ *with two different comparators:* $y^1 = y(\boldsymbol{x}, \boldsymbol{z}^1), y^2 = y(\boldsymbol{x}, \boldsymbol{z}^2)$ *(differences* $\boldsymbol{\delta}^1$ *and* $\boldsymbol{\delta}^2$*). Two conditions must be satisfied: (1) the evaluations are not equal* ($y^1 \neq y^2$)*, and (2) the second comparison differs on a superset of attributes* ($\boldsymbol{\delta}^2 > \boldsymbol{\delta}^1$)*.*

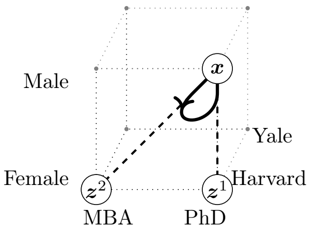
Corollary 4. *A convex scissor implies at least one nonzero implicit preference:*

$y^2 > y^1$ *(i) favoring* $\boldsymbol{x}$*'s realization of an attribute that it does not share with* $\boldsymbol{z}^1$*,*
(ii) disfavoring $\boldsymbol{x}$*'s realization of an attribute that it shares with* $\boldsymbol{z}^2$*, or*
(iii) with unrestricted sign on any other attribute.

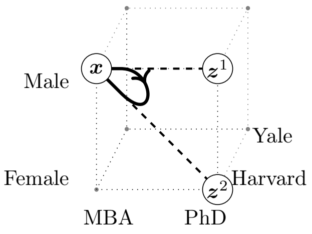
$y^2 < y^1$ *(i)* disfavoring $\boldsymbol{x}$*'s realization of an attribute that it does not share with* $\boldsymbol{z}^1$*,*
(ii) favoring $\boldsymbol{x}$*'s realization of an attribute that it shares with* $\boldsymbol{z}^2$*, or*
(iii) with unrestricted sign on any other attribute.

Defining the variable $\Upsilon = sgn(y^2 - y^1) \in \{-1, 1\}$*, we can write:*

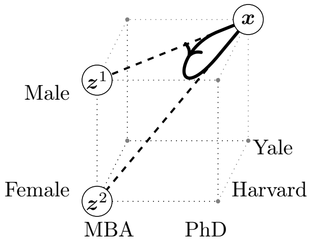
$$\bigvee_{i: x_i \neq z_i^1} (x_i \kappa_i \Upsilon = 1) \vee \bigvee_{i: x_i = z_i^2} (x_i \kappa_i \Upsilon = -1) \vee \bigvee_{i: z_i^1 \neq z_i^2} (\kappa_i \neq 0).$$



$y^2 > y^1$
$(\kappa_1 \neq 0) \vee (\kappa_2 > 0) \vee (\kappa_3 > 0)$
±PhD $\vee$+Male $\vee$+Yale.



$y^2 < y^1$
$(\kappa_1 > 0) \vee (\kappa_2 \neq 0) \vee (\kappa_3 < 0)$
+PhD $\vee$±Male $\vee$−Yale.



$y^2 < y^1$
$(\kappa_1 < 0) \vee (\kappa_2 \neq 0) \vee (\kappa_3 < 0)$
−PhD $\vee$±Male $\vee$−Yale.

Intuitively, the shift of comparison from $\boldsymbol{z}^1$ to $\boldsymbol{z}^2$ changes opacity for every attribute. Those that are shared in both comparisons become less dilute (as the set of shared attributes shrinks), so evaluation becomes less sensitive to implicit preferences on these attributes. Those that are non-shared in both comparisons become *more* dilute (as the set of non-shared attributes grows) so their implicit preferences have more influence. But those that were shared in $\{\boldsymbol{x}, \boldsymbol{z}^1\}$ but are not shared in $\{\boldsymbol{x}, \boldsymbol{z}^2\}$ are not restricted by Assumption 4 (Dilution) so we cannot sign their implicit preferences. Assumption 5 (Dominance of attribute k) resolves the ambiguity.

Corollary 5 (Convex scissor with Dominance of attribute k)**.** *When Assumption 5 holds, a convex scissor implies:*

$$\bigvee_{i:x_i \neq z_i^1} (x_i \kappa_i \Upsilon = 1) \vee \bigvee_{i:x_i = z_i^2} (x_i \kappa_i \Upsilon = -1) \vee \bigvee_{i:z_i^1 \neq z_i^2} (x_i \kappa_i \Upsilon = -\Theta),$$

where Θ equals 1 when k is shared (opacity is higher for shared attributes), and -1 when k is non-shared (opacity is higher for non-shared attributes).

Definition 9 (Parallel convex scissors)**.** *A pair of parallel convex scissors is a dataset consisting of two convex scissors, $y(\boldsymbol{x}, \boldsymbol{z}^1) \neq y(\boldsymbol{x}, \boldsymbol{z}^2)$ and $y(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^1) \neq y(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^2)$, $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$. We label the differences $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \bar{\boldsymbol{\delta}}^1, \bar{\boldsymbol{\delta}}^2$, the evaluation values $y^1, y^2, \bar{y}^1, \bar{y}^2$, and the signs of changes in evaluation values $\Upsilon = sgn(y^2 - y^2)$ and $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^2)$.*

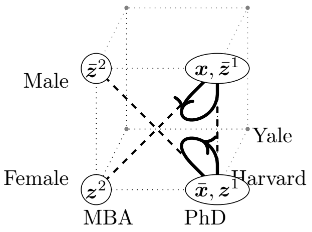
Two conditions must be satisfied: (1) identical or opposing signed *differences on $\{\boldsymbol{x}, \boldsymbol{z}^1\}$ and $\{\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^1\}$ (i.e., either $\boldsymbol{x} - \boldsymbol{z}^1 = \bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}^1$ or $\boldsymbol{x} - \boldsymbol{z}^1 = -(\bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}^1)$); and (2) identical* absolute *differences on $\{\boldsymbol{x}, \boldsymbol{z}^2\}$ and $\{\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^2\}$ (i.e., $\boldsymbol{\delta}^2 = \bar{\boldsymbol{\delta}}^2$).*[11]

Corollary 6. *A pair of parallel convex scissors imply at least one nonzero implicit preference. There are many cases, which depend on the relationships between $\boldsymbol{x}, \bar{\boldsymbol{x}}, \Upsilon$, and $\bar{\Upsilon}$. The cases*

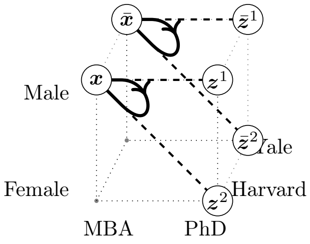
[11] We also assume that the only information derived from the evaluations is the ranking of y^1, y^2 and the ranking of $\bar{y}^1, \bar{y}^2$, i.e. we either do not have, or do not exploit, the ranking of evaluations between the scissors. In principle such information could be used in combination with functional form assumptions to extract additional information, but we do not model this for sake of brevity.

are summarized in the following disjunction:

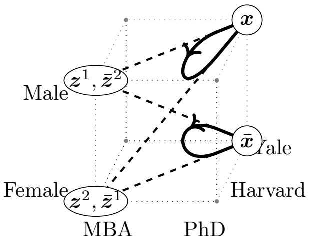
$$\bigvee_{i:x_i\neq z_i^1}\big(\kappa_i(x_i\Upsilon+\bar{x}_i\bar{\Upsilon})=1\big)\vee\bigvee_{i:x_i=z_i^2}\big(\kappa_i(x_i\Upsilon+\bar{x}_i\bar{\Upsilon})<0\big)\vee\bigvee_{i:z_i^1\neq z_i^2}\big(\kappa_i(x_i\Upsilon+\bar{x}_i\bar{\Upsilon})\neq 0\big)\,.$$



$y^2 > y^1$, $\bar{y}^2 < \bar{y}^1$
$(\kappa_2 > 0)$
+Male.



$y^2 > y^1$, $\bar{y}^2 > \bar{y}^1$
$(\kappa_1 > 0) \vee (\kappa_2 \neq 0)$
+PhD ∨±Male.



$y^2 < y^1$, $\bar{y}^2 > \bar{y}^1$
$(\kappa_2 \neq 0)$
±Male.

Just like the pair of parallel right triangles, parallel convex scissors can provide a refinement on the implications of their constituent scissors. This occurs when $x_i\Upsilon = -\bar{x}_i\bar{\Upsilon}$, in which the term associated with that attribute equals zero and is dropped from the disjunction. Intuitively, the observed behavior cannot be explained by that attribute if evaluation moves in contradictory directions in the two scissors.

Often we can eliminate all but one attribute, in which case we may uniquely identify that attribute's implicit preference. The third example illustrates a case where we uniquely identify which attribute must have an implicit preference, but not its sign, because Assumption 4 does not restrict the opacity change for this attribute. Assumption 5 resolves the ambiguity.

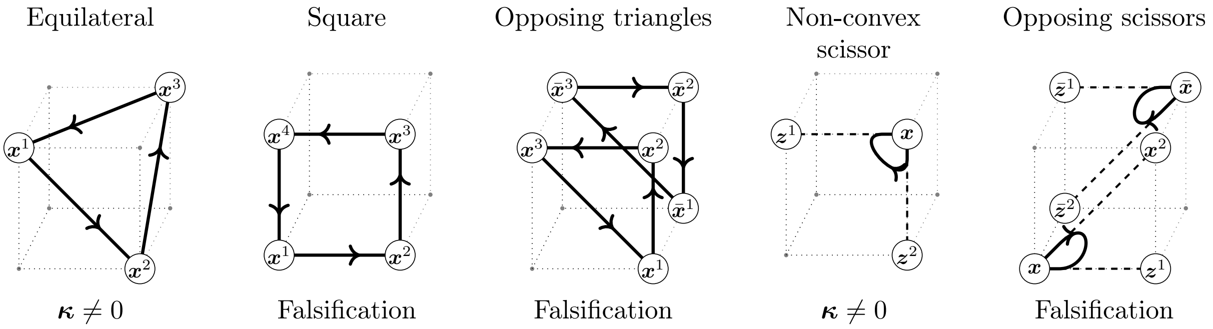
Corollary 7 (Parallel convex scissors with Dominance of attribute k)**.** *When Assumption 5 holds, a pair of parallel convex scissors implies:*

$$\bigvee_{i:x_i\neq z_i^1}\big(\kappa_i(x_i\Upsilon+\bar{x}_i\bar{\Upsilon})=1\big)\vee\bigvee_{i:x_i=z_i^2}\big(\kappa_i(x_i\Upsilon+\bar{x}_i\bar{\Upsilon})<0\big)\vee\bigvee_{i:z_i^1\neq z_i^2}\big(\kappa_i(x_i\Upsilon+\bar{x}_i\bar{\Upsilon})=-\Theta\big)\,,$$

where Θ *equals* 1 *when* k *is shared (opacity is higher for shared attributes), and* -1 *when* k *is non-shared (opacity is higher for non-shared attributes).*

Suppose we assume that the dominating attribute k is shared, so opacity is higher for shared attributes ($\Theta = 1$). In our second parallel convex scissors example, this implies $(\kappa_1 > 0) \vee (\kappa_2 > 0)$, an implicit preference favoring PhDs, or men. In our third example it implies $\kappa_2 > 0$, an implicit preference favoring men.

Other examples Figure 4 gives five examples that either fail to identify any implicit preference, or cannot be rationalized by any $\boldsymbol{\kappa}$, falsifying the model. The falsifications have in common that for every $i, \boldsymbol{\delta}$ they have an equal number of wins and losses, meaning that their score vector equals zero. Thus, wins trivially opacity dominate losses, and vice versa, so no implicit preference can rationalize the dataset.



An **equilateral triangle** is a choice cycle over three bundles, where Dilution does not rank its comparisons. Its wins neither opacity dominate its losses, nor vice versa. We cannot rule out any $\boldsymbol{\kappa}$.

A **square** cycle is a choice cycle over four bundles, in which each choice is twinned with another with the same $\boldsymbol{\delta}$ but opposing preference. Thus its score is a vector of zeros, so cannot be rationalized.

A pair of **opposing triangles** is two right triangles with opposing signed differences on each choice. Just like the square cycle, its score vector equals zero, so cannot be rationalized.

A **non-convex scissor** is one in which Dilution does not rank its comparisons, so under Separable Implicit Preferences we cannot rule out any $\boldsymbol{\kappa}$.

A pair of **opposing scissors** is one in which $\boldsymbol{x} = -\bar{\boldsymbol{x}}$ but both scissors move in the same direction. Thus its score vector equals zero, so cannot be rationalized.

Figure 4: Examples that do not identify an implicit preference, or falsify the model

4 Foundations

We now provide models of three types of decision maker: one constrained by rules (*ceteris paribus*), one concerned for their reputation (*signaling*), and one influenced by unconscious *implicit associations.*

We can explain concisely why each model satisfies our core intuition: that the influence of implicit preferences increases as comparisons become more opaque. In the *ceteris paribus* model, the decision maker is constrained by rules that apply to certain comparisons (e.g., a male MBA versus a female MBA) but turn off as the comparison becomes more dilute (a male MBA versus a female PhD). As a result, they can express their implicit preferences more strongly as the set of differences between bundles grows. In the signaling model, the more that an attribute is mixed with others, the less an observer can infer about the decision maker's preference for that attribute, so they feel freer to express their true preferences. In the implicit associations foundation, the more an attribute is mixed with others, the less the decision maker can infer about potential unconscious influences on their preferences, so the more inclined they are to go with their gut instincts.

It will be useful to define the set of shared attributes for comparison $|\boldsymbol{x}-\boldsymbol{z}|$:

$$S^{|\boldsymbol{x}-\boldsymbol{z}|}=\{i:|x_i-z_i|=0\}.$$

Non-shared attributes are those not in S. We suppress the superscript unless needed.

Proof strategy. We provide conditions under which each model satisfies Separable Implicit Preferences. Each can be expressed as a comparative utility function satisfying Assumption 1, where $\theta_i()$ depends only on $|\boldsymbol{x}-\boldsymbol{z}|$ (Assumption 3) and is increasing as $|\boldsymbol{x}-\boldsymbol{z}|$ becomes more dilute (Assumptions 2 and 4).

In each foundation $\theta_i(.)$ takes a particular form that depends on i's status:

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|)=\begin{cases}\theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & ,i\in S\\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & ,i\notin S.\end{cases}$$

We need to show that $\theta_i()$ is weakly increasing as i becomes more dilute. By definition of dilution, $|\boldsymbol{x}'-\boldsymbol{z}'|\sqsupseteq_i|\boldsymbol{x}-\boldsymbol{z}|$ when (a) i does not change status (either $(i\in S^{|\boldsymbol{x}'-\boldsymbol{z}'|})\wedge(i\in S^{|\boldsymbol{x}-\boldsymbol{z}|})$ or $(i\notin S^{|\boldsymbol{x}'-\boldsymbol{z}'|})\wedge(i\notin S^{|\boldsymbol{x}-\boldsymbol{z}|})$; and (b) the set of attributes that share status with i grows ($\{j:|\boldsymbol{x}_j'-\boldsymbol{z}_j'|=|\boldsymbol{x}_i'-\boldsymbol{z}_i'|\}\supseteq\{j:|\boldsymbol{x}_j-\boldsymbol{z}_j|=|\boldsymbol{x}_i-\boldsymbol{z}_i|\}$).

Part (a) implies that we can study the properties of θ_i^S and θ_i^N separately, since i does not change status in a given dilution. Part (b) implies we need to show that $\theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|)$ weakly increases as the set of shared attributes grows (in a superset sense), and that $\theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|)$ weakly increases as the set of non-shared attributes grows.

4.1 *Ceteris Paribus* Decision Maker

Suppose our hiring manager normally chooses whichever candidate they prefer, except when comparing a male candidate to an otherwise identical female candidate, in which case they are compelled to hire the woman. We state a general model of *ceteris paribus* decision makers who are constrained by rules that apply "all else equal," but otherwise maximize menu-independent utility. Rules can be interpreted as internal to the decision maker (e.g. a moral obligation) or external (e.g. a bureaucratic rule).

We will also allow multiple rules which can compound or counteract one another, in which case "all else equal" is taken to mean when all *non-rule-governed* attributes are equal. Suppose a manager is instructed to both (1) prefer female candidates all else equal, (2) prefer

Black candidates all else equal. We will assume that the rules combine such that they must choose a Black woman over a white man (otherwise equal), but when choosing between a white woman and a Black man the decision is governed by whichever rule has more force.[12]

Definition 10. *A **ceteris paribus utility function** has the form:*

$$u^{CP}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{g(\boldsymbol{x})}_{\substack{\textit{menu-independent}\\ \textit{utility}}} + \sum_{i \notin S} x_i \underbrace{\lambda_i}_{\substack{\textit{bonus or}\\ \textit{penalty}}} \underbrace{\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S)\}}_{\substack{\textit{=1 iff all non-rule-governed}\\ \textit{attributes are shared}}},$$

for some $g : \{-1,1\}^n \to \mathbb{R}$, and $\boldsymbol{\lambda} \in \mathbb{R}^n$.

When $\lambda_i \neq 0$ we say attribute i is governed by a rule. Thus the bonus/penalty λ_i is applied to a bundle if and only if (a) attribute i is non-shared ($i \notin S$); and (b) every attribute that is not governed by a rule ($\lambda_j = 0$) is shared ($j \in S$).

Applied to choice, λ_i is a bonus/penalty for choosing one bundle over another. Rules could demand hiring a Black candidate, booking the cheapest flight, or ordering a low-calorie meal. If $\lambda = \infty$ the rule is inviolable. Applied to evaluation, λ_i is a bonus/penalty applied to reported values. For example, someone might give women lower scores except when they are compared to an otherwise-identical man.

$u^{CP}(\boldsymbol{x},\boldsymbol{z})$ can be rearranged to satisfy Assumption 1:

$$u^{CP}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{(-sgn(\lambda_i))}_{\kappa_i} \theta_i(|\boldsymbol{x}-\boldsymbol{z}|)$$

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i| & , i \in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i|(1 - \mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S)\}) & , i \notin S. \end{cases}$$

Proposition 1. *$u^{CP}(\boldsymbol{x},\boldsymbol{z})$ satisfies Separable Implicit Preferences.*

For interpretation, consider our hiring manager example. When the candidates differ only on gender, opacity is low (the rule is applied) decreasing the utility of the male candidate. When the candidates differ in other attributes as well, opacity increases (the rule is turned off), increasing the utility of the male candidate increases. Hence, the rule manifests as an implicit preference favoring men.

[12] There are other possible interpretations of interactions between "all else equal" rules. A literal interpretation would say rules only apply when the bundles differ on exactly one attribute. This satisfies separable implicit preferences is counterintuitive: in our example, the decision maker would be allowed to choose a white male over a Black female. Another interpretation is that rules only apply when they all agree: Black women are always chosen over white men, but no rule applies when comparing a Black man to a white woman. This setup will not satisfy Assumption 3, because θ then depends on the signed differences between bundles $(\boldsymbol{x}-\boldsymbol{z})$.

4.2 Signaling Decision Maker

Suppose the decision maker holds intrinsic values over attributes, but also has reputational preferences. They care about the beliefs that some other person—perhaps their own future self—holds over those intrinsic values. We represent their intrinsic values as:

$$g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i$$

where $g(\boldsymbol{x})$ is assumed to be common knowledge, while w_i terms ("weights") are the decision maker's private information. We assume the observer holds mean-zero, independent Normal priors over the w_is, and forms posteriors based on the decision maker's actions.

To be more precise we must specify the information structure of the game, and that structure differs between choice and evaluation. We will thus describe separate signaling models for each type of behavior. We assume throughout that the bundles $\boldsymbol{x}$ and $\boldsymbol{z}$ are chosen by Nature and are common knowledge (i.e., we abstract from strategic choice over choice sets).

4.2.1 Choice

When the decision maker chooses $\boldsymbol{x}$ over $\boldsymbol{z}$ the observer will update their beliefs $\hat{w}_i$ about the decision maker's weights on attributes where $\boldsymbol{x}$ and $\boldsymbol{z}$ differ. The core intuition is that when the bundles differ on a superset of attributes, the observer updates less about each w_i, so the decision maker's signaling incentives weaken.

We make two strong assumptions, which amount to the observer expecting the decision maker to be indifferent *ex ante*. First, the observer's priors over all intrinsic values have identical mean, which we normalize to zero: $g(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathcal{X}$. Second, we assume the observer is *naïve*, meaning they are not aware of the decision maker's reputational motives.[13] For applications it may not be reasonable to assume sharp mean-zero priors, but our conclusions should be valid for quantitatively modest deviations. We discuss the relevance of mean-zero priors to applications in Section 6.

We define a comparative utility function $u^{SC}(\boldsymbol{x}, \boldsymbol{z})$, interpreted as the utility of choosing $\boldsymbol{x}$ when the observer knows the choice set was $\{\boldsymbol{x}, \boldsymbol{z}\}$. We initially assume all preferences

[13] If instead the observer had reason to believe the decision maker prefers one bundle over another then changing the comparator to make the difference more dilute can make the choice *more* informative about an attribute rather than less, violating assumption 4. For example, choosing a male PhD over a female MBA is less informative about gender preferences than choosing a male PhD over a female PhD. But choosing a male PhD over a female Nobel prize winner is *more* informative, in the sense of posteriors being farther apart. If the observer had priors over reputational motives, they would again have a prior over which bundle would be chosen. Once again more dilute comparisons are not guaranteed to be less informative about weights.

are strict, and discuss indifference at the end of the section. We also assume $\boldsymbol{x}$ and $\boldsymbol{z}$ are distinct, so there is at least one non-shared attribute.

Definition 11. *A **signaling-choice utility function** has the form:*

$$\underbrace{u^{SC}(\boldsymbol{x},\boldsymbol{z})}_{\substack{\textit{utility of}\\ \textit{choosing } \boldsymbol{x}\\ \textit{from } \{\boldsymbol{x},\boldsymbol{z}\}}} = \underbrace{\sum_{i=1}^{n} x_i w_i}_{\substack{\textit{intrinsic}\\ \textit{value}}} + \sum_{i=1}^{n} \underbrace{\lambda_i}_{\substack{\textit{reputational}\\ \textit{preference}\\ \textit{for attribute } i}} \cdot \underbrace{E\left[w_i \middle| \sum_{i=1}^{n} x_i w_i > \sum_{i=1}^{n} z_i w_i\right]}_{\substack{\textit{observer's naïve posteriors}\\ \textit{over weights when } \boldsymbol{x} \textit{ is chosen}}},$$

for some $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$ (observer's priors over weights).

$\boldsymbol{\lambda}$ captures the decision maker's utility of shifting the observer's posteriors over weights. We can derive an explicit solution for the observer's posterior:

Lemma 1. *Suppose a naïve observer sees the decision maker choose $\boldsymbol{x}$ from $\{\boldsymbol{x},\boldsymbol{z}\}$, $\boldsymbol{x} \neq \boldsymbol{z}$. Their posterior over weight w_i can be written as:*

$$E\left[w_i \middle| \sum_{i=1}^{n} x_i w_i > \sum_{i=1}^{n} z_i w_i\right] = \mathbf{1}\{i \notin S\} \frac{x_i \sigma_i^2}{\sqrt{\sum_{j \notin S} \sigma_j^2}} \frac{\phi(0)}{1-\Phi(0)},$$

where ϕ and Φ are the standard Normal density and cumulative density functions.

Three things are worth noting. First, the observer divides attribution for the choice among the weights w_i on non-shared attributes, attributing more to those with larger variance σ_i^2. Second, the magnitude of the belief change on a given non-shared attribute i is decreasing as the set of non-shared attributes grows, i.e. as the comparison becomes more dilute with respect to i. Third, they do not update at all about weights on shared attributes, since choice is uninformative about those weights. Thus there is no reputational effect for shared attributes.

Using the lemma, we can rearrange u^{SC} to satisfy Assumption 1:

$$u^{SC}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{\sum_{i=1}^{n} x_i \left(w_i + \lambda_i \sigma_i \frac{\phi(0)}{1-\Phi(0)}\right)}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{(-sgn(\lambda_i))}_{\kappa_i} \theta_i(|\boldsymbol{x}-\boldsymbol{z}|)$$

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i|\sigma_i \frac{\phi(0)}{1-\Phi(0)} & , i \in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i|\sigma_i \left(1 - \frac{\sigma_i}{\sqrt{\sum_{j \notin S}\sigma_j^2}}\right) \frac{\phi(0)}{1-\Phi(0)} & , i \notin S. \end{cases}$$

Proposition 2. *$u^{SC}(\boldsymbol{x},\boldsymbol{z})$ satisfies Separable Implicit Preferences.*

Our setup assumes the decision maker can only express strict preferences. It is possible to show that a decision maker would choose to express indifference, with its consequent reputational effects, only if they received equal utility from expressing indifference and expressing either of the two strict preferences, i.e. $u^{SC}(\boldsymbol{x}, \boldsymbol{z}) = u^{SC}(\boldsymbol{z}, \boldsymbol{x})$. Thus the function we derived for the 2-action world correctly predicts behaviour in a 3-action world, so the model can be applied to data containing indifferences.[14]

For interpretation, consider the hiring manager that prefers men but wants the observer to believe they prefer women. When candidates differ on few attributes, the observer infers a lot about their gender preferences from their choice. As additional attributes vary, the observer updates less about gender, lowering the reputational cost of hiring a man. Note that the implicit preference κ_i has the opposite sign to its associated signaling motive λ_i: a motive to signal a preference for women manifests as an implicit preference favoring men.

4.2.2 Evaluation

In evaluation we assume the decision maker considers two bundles, $\boldsymbol{x}$ and $\boldsymbol{z}$, and assigns each a value, y^x and y^z, with a quadratic cost of inaccuracy. An observer then makes inferences about the decision maker's weights w_i. Unlike the choice setting, we do not need to assume that the observer has constant priors over the weights, nor that they are naïve.

We will first define a signaling evaluation function, $u^{SE}(\boldsymbol{x}, \boldsymbol{z})$, then show that it corresponds to an equilibrium strategy in a signaling game, and finally that it satisfies Separable Implicit Preferences.

Definition 12. *A* ***signaling evaluation utility function*** *is:*

$$u^{SE}(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i + \sum_{i=1}^{n} x_i \lambda_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\} \sum_{j \in S} \sigma_j^2 + \mathbf{1}\{i \notin S\} \sum_{j \notin S} \sigma_j^2}$$

for some $g : \{-1, 1\}^n \to \mathbb{R}$, $\boldsymbol{w}, \boldsymbol{\lambda} \in \mathbb{R}^n$, $\boldsymbol{\sigma} \in \mathbb{R}^n_+$.

$\boldsymbol{\lambda}$ captures the decision maker's utility of shifting the observer's posteriors over weights. The final term represents an adjustment that the decision maker applies to her evaluations in order to influence the observer's beliefs. The adjustment to attribute i's value is proportional to the observer's uncertainty about the intrinsic weight on i (σ_i^2), and inversely proportional to the total uncertainty about the attributes with which i shares status.

Lemma 2. *Reporting the value of* $y^x = u^{SE}(\boldsymbol{x}, \boldsymbol{z}), y^z = u^{SE}(\boldsymbol{z}, \boldsymbol{x})$*, is an optimal strategy in a pure-strategy Perfect Bayes Equilibrium of a signaling game in which:*

[14] A derivation is available on request.

1. *Player 1 first chooses y^x and y^z to maximize*

$$U^1 = \underbrace{-\frac{1}{2}\Big(y^x - g(\boldsymbol{x}) - \sum_{i=1}^{n} w_i x_i\Big)^2 - \frac{1}{2}\Big(y^z - g(\boldsymbol{z}) - \sum_{i=1}^{n} w_i z_i\Big)^2}_{\text{quadratic loss from inaccuracy}} + \underbrace{\sum_{i=1}^{n} \lambda_i \hat{w}_i(y^x, y^z)}_{\text{reputational gain}}.$$

2. *Player 2 observes y^x, y^z and chooses $\hat{\boldsymbol{w}}$ to maximize*

$$U^2 = -E\left[\sum_{i=1}^{n} (\hat{w}_i - w_i)^2 \middle| y^x, y^z\right],$$

with $g(\cdot)$ and $\boldsymbol{\lambda}$ common knowledge, and priors $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$.

We stated at the beginning that we did not need to assume naïveté, and solved the model assuming full sophistication. However it is worth noting that player 1's best response is independent of the observer's beliefs about $\boldsymbol{\lambda}$ and so our solution continues to hold if the observer is naïve or partially naïve.[15]

Remark 1. *A strategy of reporting the value of a signaling-evaluation utility function remains optimal if the observer incorrectly believes the signaling motive is $\tilde{\boldsymbol{\lambda}} \neq \boldsymbol{\lambda}$.*

Sophistication corresponds to $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}$ while full nav̈eté corresponds to $\tilde{\boldsymbol{\lambda}} = 0$.

We can write $u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ in a form that satisfies Assumption 1:

$$u^{SE}(\boldsymbol{x}, \boldsymbol{z}) = \underbrace{g(\boldsymbol{x}) + \sum_{i=1}^{n} (w_i + \lambda_i) x_i}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{(-sgn(\lambda_i))}_{\kappa_i} \theta_i(|\boldsymbol{x} - \boldsymbol{z}|)$$

$$\theta_i(|\boldsymbol{x} - \boldsymbol{z}|) = \begin{cases} \theta_i^S(|\boldsymbol{x} - \boldsymbol{z}|) & = |\lambda_i| \left(1 - \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2}\right) \quad, i \in S \\ \theta_i^N(|\boldsymbol{x} - \boldsymbol{z}|) & = |\lambda_i| \left(1 - \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2}\right) \quad, i \notin S. \end{cases}$$

Proposition 3. *$u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ satisfies Separable Implicit Preferences.*

The intuition behind how signaling motives manifest as implicit preferences in evaluation is very similar to the choice example, with the exception that the observer now updates about both shared and non-shared attributes, because they observe distinct signals about both bundles' values rather than just their ranking.

[15] The derivatives $\partial \hat{w}_i / \partial y^x$ and $\partial \hat{w}_i / \partial y^z$ do not depend on player 2's belief about $\boldsymbol{\lambda}$, so replacing $\boldsymbol{\lambda}$ with some alternative $\tilde{\boldsymbol{\lambda}}$ in $\hat{w}_i$ does not affect player 1's optimal strategy. The mechanism is that player 2's belief about $\boldsymbol{\lambda}$ determines by how much they expect y^x to over- or understate $\boldsymbol{x}$'s true intrinsic value on average, but not how they respond to marginal *changes* in y^x. Similarly, it does not affect player 1's behavior if player 2's mean priors over intrinsic values are wrong, $\tilde{g}(\boldsymbol{x}) \neq g(\boldsymbol{x})$.

4.3 Implicit Associations Decision Maker

Finally we describe a decision maker made up of two agents, each with private information relevant to the value of a bundle:[16]

$$\underbrace{f(\boldsymbol{x})}_{\substack{\text{true value}\\ \text{of bundle } \boldsymbol{x}}} = \underbrace{g(\boldsymbol{x})}_{\substack{\text{known}\\ \text{by both}}} + \sum_{i=1}^{n} \underbrace{x_i}_{\substack{\text{known}\\ \text{by both}}} \cdot \underbrace{\lambda_i}_{\substack{\text{known}\\ \text{by first}\\ \text{agent}}} \cdot \underbrace{\pi_i}_{\substack{\text{known}\\ \text{by second}\\ \text{agent}}} .$$

The first agent can be thought of as the pre-conscious brain, drawing on knowledge of "associations" ($\boldsymbol{\lambda} \in \mathbb{R}^n$) between each attribute and true value, and the second agent can be thought of as the conscious brain, which has access to "adjustments" ($\boldsymbol{\pi} \in \mathbb{R}^n_+$), high-level contextual information used to adjust the value of each association.

Sequencing is as follows. The first agent calculates expected values for bundles $\boldsymbol{x}$ and $\boldsymbol{z}$ ($E[f(\boldsymbol{x})|\boldsymbol{\lambda}]$ and $E[f(\boldsymbol{z})|\boldsymbol{\lambda}]$). The second agent then makes decisions taking into account the first agent's estimates, plus its own private information ($\boldsymbol{\pi}$), but without access to the underlying associations ($\boldsymbol{\lambda}$). The theory predicts that the second agent's estimate of $\boldsymbol{x}$'s value will be affected by a comparator $\boldsymbol{z}$ insofar as the comparison is informative about associations, $\boldsymbol{\lambda}$.

The core idea is that associations are generally considered informative (otherwise the second agent would ignore the first agent's estimates). However, the second agent has access to contextual information that leads her to adjust the first agent's estimates. Her ability to apply these adjustments depends on the degree to which she can separately distinguish the influence of each association. For example, the decision maker might have an association with gender, $\lambda_i \neq 0$, e.g. an unconscious positive attitude toward men. However, the conscious brain has reason to believe that in the current setting any such association is normatively irrelevant ($\pi_i = 0$). We expect such a decision maker to exhibit a pro-male bias that decreases as the comparison becomes less opaque.

Definition 13. *An* ***implicit associations utility function*** *has the form:*

$$u^{IA}(\boldsymbol{x}, \boldsymbol{z}) = E[f(\boldsymbol{x})|\boldsymbol{\pi}, \hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})],$$

[16] This model is based on Cunningham (2014), which discusses more generally conditions under which sequential aggregation of information will be efficient.

with

$$\begin{aligned} &\hat{f}(.) = E[f(.)|\boldsymbol{\lambda}] && \textit{(1st agent's estimate of } f(.)\textit{)} \\ &\pi_i \in \mathbb{R}_+ \;\&\; E[\pi_i] = 1 && \textit{(1st agent's priors)} \\ &\boldsymbol{\lambda} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2)) && \textit{(2nd agent's priors)} \\ &\boldsymbol{\pi} \perp\!\!\!\perp \boldsymbol{\lambda} && \textit{(independence of priors).} \end{aligned}$$

Next, we show that the utility function takes the following simple form:

Lemma 3. *An implicit associations utility function can be written:*

$$u^{IA}(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i \bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|)$$

$$\bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|) = \begin{cases} \frac{\sum_{j \in S} \pi_j \sigma_j^2}{\sum_{j \in S} \sigma_j^2} & , i \in S \\ \frac{\sum_{j \notin S} \pi_j \sigma_j^2}{\sum_{j \notin S} \sigma_j^2} & , i \notin S \end{cases}$$

u^{IA} represents the second agent's best guess at the value of the bundle. She would ideally like to weight each attribute by $\lambda_i \pi_i$, but instead weights them by $\lambda_i \bar{\pi}_i$, where $\bar{\pi}_i$ represents the variance-weighted average of π among the group of attributes with which attribute i shares status.

The implicit associations model differs from the other models. The sensitivity of utility with respect to a given attribute will be proportional to the average π_i among the attributes with which that attribute is grouped. As a consequence implicit preferences are relative, rather than absolute: if attribute 1 has an intermediate value of π_i then it sensitivity could increase when diluted with one attribute but decrease when diluted with another. This is inconsistent with Assumption 4.

The foundation is consistent with Separable Implicit Preferences in two special cases: (i) when there are exactly two attributes, $n = 2$; (ii) when an unexpected realization of information occurs for at most one attribute (at most one i has either $\lambda_i \neq 0$ or $\pi_i \neq 1$). We adopt the second assumption for the remainder of the section.[17]

If we assume that at most one attribute has a non-zero implicit association ($\lambda_i \neq 0$) or non-unity adjustment factor ($\pi_i \neq 1$), then u^{IA} can be written as:

[17] There is a certain artificiality: in effect the econometrician has information that is unobserved by the agents (the realizations of λ_i and π_i for the $n - 1$ remaining attributes).

$$u^{IA}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i\lambda_i\pi_i}_{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i \underbrace{sgn(\lambda_i(1-\pi_i))}_{\kappa_i} \theta_i(|\boldsymbol{x}-\boldsymbol{z}|)$$

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i(1-\pi_i)| \left(1 - \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\right) \quad , i \in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i(1-\pi_i)| \left(1 - \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\right) \quad , i \notin S. \end{cases}$$

Proposition 4. *$u^{IA}(\boldsymbol{x},\boldsymbol{z})$ satisfies Separable Implicit Preferences if at most one attribute has a non-zero implicit association and/or non-unitary adjustment factor:*

$$\sum_{i=1}^{n} \mathbf{1}\{\lambda_i \neq 0 \ \ or \ \pi_i \neq 1\} \leq 1.$$

For intuition of how implicit associations can be interpreted as implicit preferences consider the hiring manager that has a positive association with male candidates (via λ_i), but believes gender is normatively irrelevant ($\pi_i = 0$). When the candidates differ only on gender, opacity is low ($\theta_i = 0$), as the second agent can directly detect and override the influence of λ_i. When gender is diluted, the influence of gender is mixed with other possible associations. As the second agent believes some of those associations might contain valuable information, they do not ignore them entirely. Thus λ_i influences their decision, increasing the utility of the man. Hence a positive association with men manifests as an implicit preference favoring men.[18]

There is only an implicit preference if both $\lambda_i \neq 0$ *and* $\pi_i \neq 1$. If $\lambda_i = 0$ the first agent's forecasts are independent of the comparator, so comparisons do not influence utility. If $\pi_i = 1$ the second agent does not want to adjust the first agent's forecasts, so their influence is independent of opacity.

4.4 Dominance of attribute k

When analyzing evaluation data we introduce Assumption 5, which says that a special attribute "k" increases opacity for all attributes that share its status. Our final proposition provides sufficient conditions for this assumption to hold in all foundations.

Proposition 5. *The* ceteris paribus *decision maker of Proposition 1 satisfies Assumption 5 if k is not governed by a rule. The* signaling evaluation *decision maker of Proposition*

[18] Note that the sign of the implicit preference depends on $\lambda_i(1-\pi_i)$. If $\pi_i > 1$, the second agent wants to *amplify* their implicit associations (in a sense, they think the first agent is too conservative). This generates an implicit preference with the opposite sign to λ_i. In our example, this would generate a strong positive preference for men when the candidates differ only on gender, weakening as gender is mixed with other attributes, and thus manifesting as an implicit preference favoring women.

3 satisfies Assumption 5 if $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$. The implicit associations *decision maker of Proposition 4 satisfies Assumption 5 if $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$.*

5 Related Theories

Our identification of implicit preferences relies on inconsistencies in choice and in evaluation. However inconsistencies could occur for other reasons, so in this section we discuss three classes of models of anomaly, and argue that each is unlikely or unable to produce the specific patterns in choice and evaluation that we associate with implicit preferences.

Contingent weighting. Models of contingent weighting in multi-attribute choice, like our theory, assume that preferences depend on the choice set.[19] However most existing theories rely on a very different intuition: they assume that the sensitivity to a given attribute depends on the observed distribution over that attribute. In contrast, our model assumes that sensitivity depends on what other attributes share status with the attribute of interest. None of the recently published contingent-weighting models is consistent with a figure-8 intransitivity.[20]

A similar point applies to the literature on comparing joint and separate evaluation of outcomes: Hsee et al. (1999) give many examples. Most of these studies find that people are more sensitive to an attribute when presented jointly (two bundles simultaneously) than separately (one at a time). They argue that this increased sensitivity is a general feature of joint evaluation, called “evaluability.”[21] Again, this is a quite different principle to that

[19] For example in Kőszegi and Szeidl (2012) sensitivity is positively related to the range of values on an attribute, in Bushong et al. (2020) it is negatively related to the range, in Cunningham (2013) it is negatively related to the average, and in Bordalo et al. (2013) it is (roughly) negatively related to the proportional range (range divided by the average).

[20] Formally, suppose the utility function is entirely separable in each attribute, in the sense that it can be written as,

$$u(x, A) = \sum_i u_i(x_i, \{a_i^j\}_{j=1}^m),$$

where a_i^j is the ith attribute of the jth element of the choice set, A, then a figure-8 intransitivity could never occur because - using the gender example - the marginal distribution of the gender attribute remains the same in all four choice sets, thus the difference in attribute-utility (u_i) between “Male” and “Female” must remain the same. The two diagonal choice-sets must evoke the same utility function, because they have the same marginal distributions, and that utility function prefers Male to Female, all else equal. But this contradicts the choice observed in the vertical choice sets (where Female is chosen over Male). Separability holds for all the models discussed above except Bordalo et al. (2013), but that model cannot generate intransitive cycles in binary choices with two attributes.

[21] For example subjects were found to state a higher WTP for a dictionary with 10,000 entries when it was evaluated alone, than when it was evaluated alongside a dictionary with 20,000 entries and a torn cover. Kahneman and Frederick (2005) discuss a similar phenomenon: that subjects are generally more sensitive to changes in within-subjects experiments than in between-subjects experiments. The theory is further

used in this paper. Hsee's mechanism could not generate a figure-8 cycle, by an analogous argument to footnote 20. See Cunningham (2013) for a Bayesian rationalization of increased sensitivity in joint evaluation.

Inference from the choice set. We assume that the attribute values of one bundle are uninformative about the value of other bundles. If not, in principle, any pattern of choice could be rationalized. The relevant question is what types of prior beliefs could generate the patterns we observe, and whether those beliefs seem realistic. Consider our leading example of gender bias in hiring. These decisions could be rationalized by a hiring manager who (1) prefers women to men, all else equal; but (2) believes that men have better qualifications. Thus in the diagonal choice sets they prefer men, not because they are men, but because they have the qualification that men have.

In most of our applications this seems unlikely to be important because we use familiar attributes, so the scope for learning from the choice set seems small. Moreover, the explanation requires that the *intrinsic* value of an attribute be opposite to its *informational* value (in this case, being male is a negative signal about the person, but a positive signal about things that covary with maleness).[22]

Inattention/Heuristics. Much of our identification comes from comparing simple choices to more complex choices in which multiple attributes vary, we may worry that inconsistencies are due to complexity variation, as in models of inattention (Sims (2003), Caplin and Martin (2014), Woodford (2012)). It is intuitive that a decision-maker could become less sensitive to an attribute in a more complex choice situation, however it would be unusual for an increase in complexity to causes the preference for an attribute to *reverse*, as necessary for the figure 8 choice pattern.[23] An exception is Cubitt et al. (2018), in whose model the decision maker puts less weight on each attribute when more attributes vary, *but* treats money separately from other attributes. That model cannot generate strict cycles over non-monetary attributes.

6 Guidance for applications

When collecting new data, we expect researchers will typically want to target the simple identification tools introduced in Section 3. We now discuss some other practical guidance

developed in Hsee and Zhang (2010).

[22] This inference-based explanation can more easily rationalize a figure-8 which has indifferences on the vertical comparisons: e.g. if the decision-maker was indifferent between men and women, all else equal, but would choose men on the diagonal based beliefs about their qualifications.

[23] A figure-8 with indifferences could come from inattention if sensitivity to an attribute goes to zero in complex choices, though we are not aware of an inattention model with this feature.

for applications of our theory, focusing on construction of the attribute space, and sampling strategy.

Multivalued attributes Some attributes might take multiple values. For example, qualification could take three values (BA/MBA/PhD). Since our theory and technique are based on binary attributes, the data need to be transformed in order to apply it. The appropriate transformation depends on the setting. Our analysis of Exley (2016)'s data needs to address the fact that one attribute (the probability of winning a lottery prize) is multivalued. We construct binary attribute spaces around each probability, and analyze them separately.

***Ambivalence* in choice data** In choice data an important consideration arises that we refer to as *ambivalence*. Choice sets should be constructed such that participants are *expected to be close to indifferent*. There are two reasons for this, a statistical one and a theoretical one. The statistical (or "calibration") motive is that it is difficult to observe intransitivities, even when they exist, if the ordering of the explicit values $v(.)$ is very strong such that the implicit preferences that exist do not manifest in a cycle. The theoretical motive is that our signaling-choice foundation relies on the observer having equal priors over the utility of both bundles.

When there are multiple non-ambivalent attributes in the dataset, one solution is to group them together so that their combination plausibly satisfies ambivalence. For instance, while a hiring manager is unlikely to be close indifferent between a candidate with a BA and one with a PhD, they might plausibly be so between a BA with work experience, versus a PhD without.

Our analysis of Exley (2016)'s data faces this issue. The basic attributes that vary in her experiment (Recipient, Prize, and Probability of winning) are unlikely to satisfy ambivalence: all else equal, we would expect payments to self to be preferred to payments to charity, and higher prizes or probabilities to lower. We therefore construct two new attributes (Attitude and Risk) by grouping payments to self with lower prizes than payments to charity, and sure payoffs with lower prizes than risky ones, in order to plausibly restore Ambivalence.

Within-subjects data The theory assumes we observe the choices or evaluations of a single decision-maker, that is, we observe within-subjects data. A concern in such a dataset is order effects: participants' later decisions may be influenced by their earlier ones. The usual experimental technique to minimize order effects is to spread decisions over time, intersperse them with "filler" tasks or questions, or in other ways make their earlier decisions less salient or harder to remember. This appears to have been successful in Exley (2016)'s experiments, in which many participants reveal within-subject inconsistencies. This could be because the large number of choices in the experiment made it difficult for them to remember earlier choices.

If order effects are a serious concern, the standard response is to collect between-subjects data in which each participant makes only a small number of choices, often just one. This raises different challenges for evaluation of choice and evaluation data.

Between-subjects choice data Establishing the presence of intransitivities in between-subjects data is challenging, because intransitivity is difficult to distinguish from underlying heterogeneity in preferences (similar to the Condorcet paradox in pairwise voting). One (strong) solution is to impose homogeneity restrictions on preferences. Alternatively, one can test for violations of the Triangle inequality (see Regenwetter et al. (2011) for extensive discussion). To establish the presence of at least one intransitive decision-maker in choice over a, b, c, we would need to observe $Pr(a \succ b) + Pr(b \succ c) + Pr(c \succ a) > 2$, i.e. the average choice probability must strictly exceed $^2/_3$. For four-element cycles the threshold increases to $^3/_4$. The challenge of finding a setting with sufficiently strong intransitive preferences to satisfy such conditions may explain why there is relatively little robust evidence of intransitive choice (Müller-Trede et al., 2015).[24]

Between-subjects evaluation data Our tools for evaluation data carry over well to between-subjects data, provided we are willing to impose some restrictions on heterogeneity and functional form. Our application to DeSante (2013) is an example of such an analysis. Suppose we observe $t = 1, \ldots, T$ iid sampled individuals' evaluations of $\boldsymbol{x}$ with comparator $\boldsymbol{z}$. We allow for heterogeneity in $v(.)$ and $\boldsymbol{\kappa}$, with population averages $\overline{v(\boldsymbol{x})}$ and $\overline{\boldsymbol{\kappa}}$, while assuming $\boldsymbol{\theta}$ is common and determined by the structure of the comparison set.[25] We also assume evaluations are affine in utility: $y(\boldsymbol{x}, \boldsymbol{z}) = a + b \times u(\boldsymbol{x}, \boldsymbol{z})$. Normalizing $a = 0, b = 1$, the average evaluation is:

$$\frac{1}{T}\sum_{t=1}^{T}\left[v_t(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \kappa_{i,t}\theta_i(|\boldsymbol{x}-\boldsymbol{z}|)\right] \xrightarrow[T\to\infty]{} \overline{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i sgn(\overline{\kappa}_i)|\overline{\kappa}_i|\theta_i(|\boldsymbol{x}-\boldsymbol{z}|).$$

This is equivalent to a "representative evaluation," replacing κ_i with $sgn(\bar{\kappa}_i)$, and defining $\overline{\theta}_i = |\overline{\kappa}_i|\theta_i$. Thus, our usual tools can identify the sign of the *average* κ_i in the population. If we find that the average κ_i is positive, we learn that at least some part of the population has

[24] As an example, the choice proportions in Snyder et al. (1979)'s experiment do not satisfy the criterion and could be explained by heterogeneous, transitive preferences.

[25] We could allow for heterogeneity of the form $\theta_{i,s} = \alpha_{i,s}\theta_i$, then we would identify $sgn(E[\kappa_i\alpha_i])$.

positive κ_i. If we also assume that implicit preferences are aligned in the population (have weakly the same sign), we learn that sign.

7 Applications

7.1 Implicit Risk and Social Preferences (Exley, 2016)

Exley (2016) studies "the use of risk as an excuse not to give." She conducts two experiments in which participants make a sequence of choices between lotteries or sure payments, where the beneficiaries can be either themselves, or charity.[26] She uses the choice data to construct certainty equivalents, such that each lottery to self, and each lottery to charity, is valued in terms of money to self, and in terms of money to charity, and tests for variation in these certainty equivalents as the trade-off between self and charity varies. The canonical pattern is one in which participants tolerate more risk when the risk favors them (high certainty equivalents), and less risk when the risk favors charity (low certainty equivalents), relative to when there is no trade-off between payoffs to self or to charity. We can think of this behavior as revealing an implicit social preference, *implicit selfishness.*

Reanalyzing Exley's dataset using our methods for choice data, we confirm this interpretation. We find that 51 percent of participants make more selfish choices as opacity about selfishness increases. Our approach also yields new insights in the form of *implicit risk preferences.* 30 percent of participants become more risk averse when opacity about risk increases, while 15 percent become more risk tolerant. Overall, 33 percent of choices give rise to cycles consistent with an implicit preference, and there is a systematic tendency toward certain types of cycle: implicit selfishness and implicit risk-aversion are more prevalent than implicit risk-seeking.

7.1.1 Data

We need to do a little work to place Exley's data in a binary attribute framework. Appendix B.3 provides a detailed description of the data structure, how it can reveal preferences on a binary attribute space, and how we can use the same assumptions as Exley's analysis to impute certain choices that are not directly observed in the data. Here we provide a brief summary.

Exley's dataset consists of an initial *normalization* choice (to figure out at roughly what exchange rate the participant is indifferent between money to self and to charity), followed

[26] In her second experiment the other beneficiary is another participant in the study, we use "charity" throughout for brevity.

by a sequence of choice lists in which each choice is between a safe payoff or a lottery that pays a single prize with probability P. There are four types of choice list, in which the recipient of the lottery and the safe payoff can vary: either both are to self, both are to charity, or one is to self and the other to charity. She repeats the exercise for a total of seven values of P: $\{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. We do not observe choices between lotteries with different prizes or probabilities, so perform separate analysis for each value of P. Thus, each participant has seven separate opportunities to reveal their implicit preferences.

We show in the Appendix how, for each P, we can represent the empirical content of the data as four binary choices over two binary attributes, which we label Social $\in$ {Generous, Selfish} and Risk $\in$ {Safe, Risky}. We plot these in Figure 5. Not every choice set on this binary space is observed in the data. Specifically, participants do not make direct choices on the horizontal edges, between sure payments to self and charity, nor between lotteries. These are needed for analysis of implicit social preferences. In her analysis, Exley uses a *linearity in payoffs* assumption to compare choices involving money to self to those involving money to charity. Given how we construct the binary attribute space, that same assumption allows us to impute the choice (Generous, Safe) $\succ$ (Selfish, Safe), marked in blue on the diagram. We do not observe datapoints that would need to impute the reverse preferences on the horizontal choice sets. See the Appendix for details.

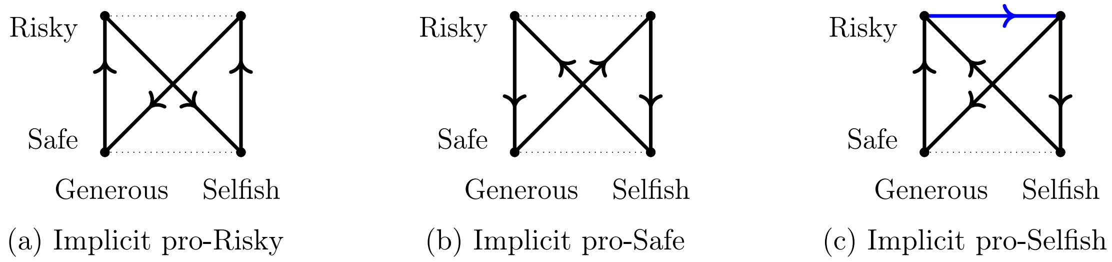
Exley excludes from most of her analysis participants whose initial normalization choices were censored or inconsistent, since their later choice lists cannot be properly calibrated. We do the same. We pool the data from both of her experiments, giving us 86 participants.

7.1.2 Observable cycles in the data

There are three patterns of choice from which we can identify a unique implicit preference. We present these in Figure 5.

Figures 5a and 5b, reveal implicit preferences over the Risk attribute. Each is a figure-8 cycle that by itself uniquely identifies an implicit preferences. Figure 5c, concerns implicit preferences over the Social attribute. The canonical pattern takes the form of two right triangles, that together uniquely identify an implicit preference for Selfish (see Section 3).[27] Note that since we do not observe the opposite horizontal choice, we cannot detect implicit Generous preferences, nor can we detect inconsistencies in implicit preferences over this attribute.

[27] Note that identification of implicit Selfish preferences depends on the imputed horizontal choice. Without it, the observed preferences could be consistent with entirely *explicit* selfish preferences. Specifically, (Selfish, Safe) $\succ$ (Selfish, Risky) $\succ$ (Generous, Risky) $\succ$ (Generous, Safe), which always ranks Selfish bundles above Generous ones. Exley's analysis depends on her linearity assumption in the same way.



(a) Implicit pro-Risky (b) Implicit pro-Safe (c) Implicit pro-Selfish

Choices marked in black are observed in the data. The choice marked in blue is imputed.

Figure 5: Exley (2016) data structure

Figure 6a presents the empirical frequencies of each of the cycles shown in Figure 5. An observation corresponds to a participant-probability pair, i.e. a set of four observed choices (plus the imputed horizontal) over which we might observe a cycle. Overall, participants exhibit one of these cycles 33 percent of time, but at different rates. Only 5 percent of choices exhibit pro-Risky cycles, 10 percent are pro-Safe, while 18 percent are pro-Selfish.[28]

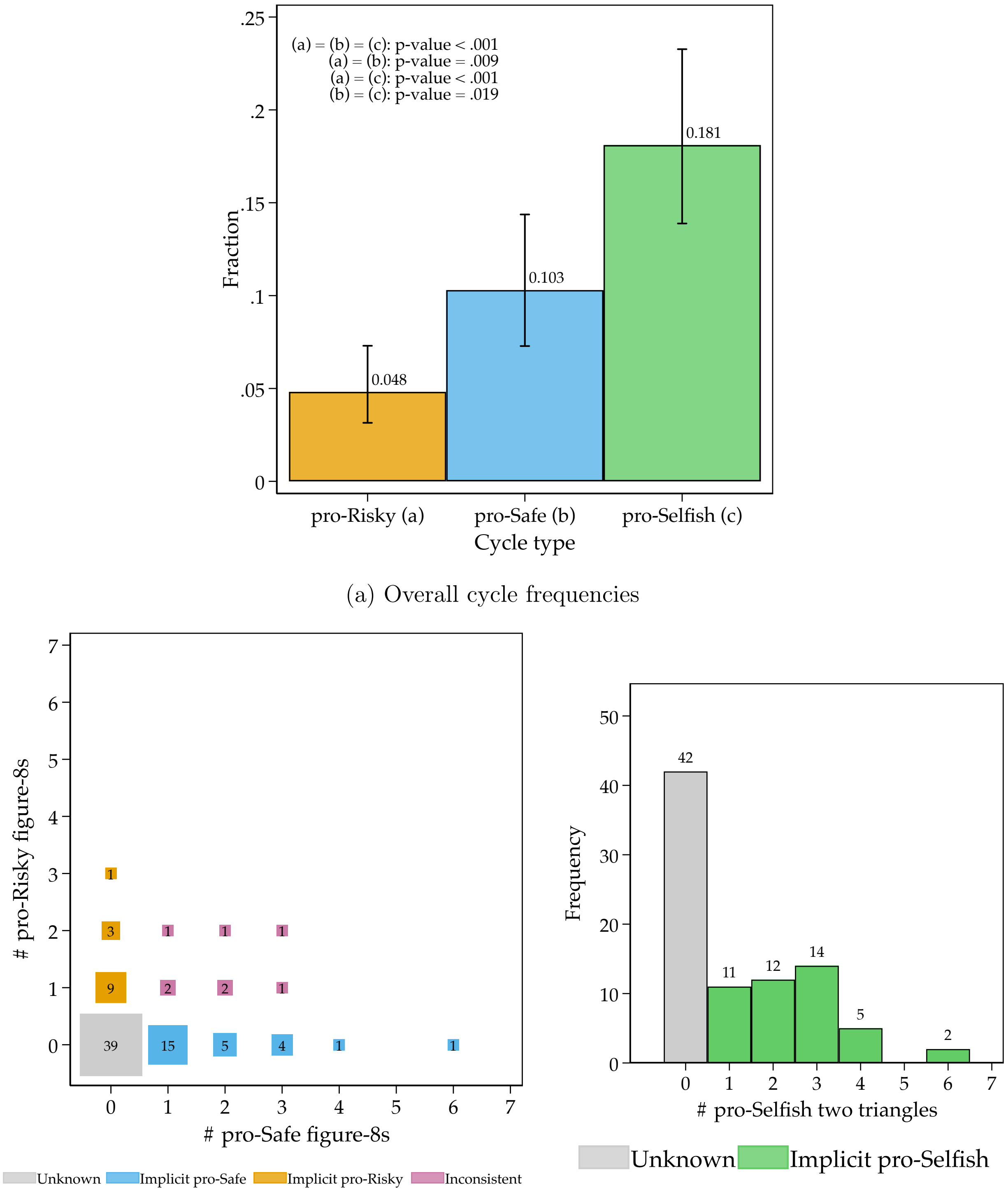
7.1.3 Classifying individuals by implicit preference type

We begin by analyzing implicit risk preferences. We classify participants into one of four categories, by counting their number of pro-Risky and pro-Safe cycles across the seven values of P. The classifications are: *Unknown* (no type (a) or (b) cycles); *Implicit pro-Risky* (at least one (a) cycle, no (b) cycles); *Implicit pro-Safe* (at least one (b) cycle, no (a) cycles); and *Inconsistent* (at least one of each). Figure 6b presents the findings, plotting the joint distribution of participant-level cycle counts. Of the 86 participants, 39 do not reveal any implicit risk preferences, 26 are implicitly pro-Safe, 13 implicitly pro-Risky, and 8 are inconsistent.

We observe many risk-related intransitive cycles, of which just over two thirds are pro-Safe. Implicit risk attitudes seem to be prevalent, and tend to be implicitly risk-averse. This could have important implications for real-world decisions. An implicitly risk-averse decision-maker might make more risk-averse choices when choosing between pension plans with different attributes (where opacity about risk is high) than she would when choosing between different variants of the same plan (where opacity is lower). That could have substantial implications for wealth at retirement.

However, relatively few participants consistently and repeatedly cycle in the same direc-

[28] An interesting benchmark is what we might observe under random choice. To calculate this benchmark, we fix the choice between (Selfish, Safe) and (Selfish, Risky), since it is used to construct the rest of the attribute space. Assuming that each other choice is 50-50 random, we would expect to see each of our three cycle patterns 12.5 percent of the time. This is similar to the average probability of a cycle, but does not explain the strong systematic tendency away from pro-Risky and toward pro-Selfish cycles.



(a) Overall cycle frequencies

(b) Implicit risk preferences

(c) Implicit social preferences

Panel (a) plots the proportion of each (mutually exclusive) cycle type at the participant-probability level. Standard errors use logit approximation, clustered at participant level. Panel (b) classifies participants according to their number of implicit pro-Risky and pro-Safe cycles. Cell size and numeric labels indicate the number of participants in each cell, colors indicate the type classification. Panel (c) classifies participants according to their number of implicit pro-Selfish cycles.

Figure 6: Type classifications in Exley (2016) data

tion. Among the consistent participants, 11 participants exhibit two or more pro-Safe cycles, while 4 exhibit two or more pro-Risky cycles. This suggests that implicit risk attitudes are weak at the individual level. We assess the overall strength of the findings in Section 7.1.4 below.

Turning to implicit Social preferences, we classify participants according to their number of pro-Selfish cycles. They can be either *Unknown* (no (c) cycles), or *Implicit pro-Selfish* (at least one (c) cycle). Figure 6c shows that just over half of the participants (44 in total) are classified as implicitly pro-Selfish. Of those, 75% exhibit this pattern twice or more. Thus implicit selfishness is more widespread, and expressed more frequently, than either pro-Safe or pro-Risky implicit preferences. However, as noted above, we cannot assess the extent of inconsistency in this preference.

In sum, like Exley, we find substantial evidence of implicit selfishness. Our analysis demonstrates that our general-purpose method can pick this up when applied to experimental choice data, and can extract new findings (implicit risk preferences) from data collected for another purpose.

7.1.4 Statistical analysis

Our analysis so far assumes behavior is deterministic, giving rise to discrete type classifications. In reality some of the heterogeneity we observe is likely a result of errors, or noise in the data.

Our first test asks whether the data are consistent with purely random behavior. Figure 6a shows that the frequency of each type of cycle is heterogeneous, and a joint test strongly rejects equality of frequencies across types of cycle ($p < .001$). Additionally, we reject equality of each pairwise difference. We observe a strong systematic tendency toward pro-Selfish cycles, and a strong tendency toward pro-Safe relative to pro-Risky cycles.

Tests 2 and 3 examine whether these general tendencies are consistent with homogeneous or "representative" implicit preferences, plus noise. We do this using permutation tests, which we describe in detail in Appendix B.3.5.

Test 2 asks whether there is systematic individual heterogeneity in behavior. The null hypothesis is that, conditional on the frequencies in Figure 6a, all participants have the same likelihood of exhibiting each type of cycle. We strongly reject this hypothesis ($p < .001$).

A key driver of the rejection in test 2 is that 17 participants do not exhibit any cycles at all, whereas we would expect only around 5 under the null. It could be that the sign of implicit preferences are homogeneous, but some individuals express them more strongly or more frequently than others. Test 3 therefore asks whether the heterogeneity we observe is consistent with *heterogeneous* cycle probabilities but *homogeneous* underlying types. The null hypothesis is that *conditional on a cycle*, the relative likelihood of each type of cycle is fixed. We find evidence against this hypothesis too, albeit somewhat weaker ($p = .065$). We conclude that while noise is likely an important factor, there is evidence of systematic and heterogeneous implicit preferences in the sample.

7.2 Implicit Racial Discrimination (DeSante, 2013)

DeSante (2013) conducted an experiment on a US representative sample, in which participants were asked to recommend state welfare payments for hypothetical applicants. The paper asks whether people reward hard work in a "color-blind manner," i.e. whether the relationship between the applicant's reported "work ethic" and the funds allocated to them is the same for black and white applicants. We will show how the experimental data can also be analyzed through the lens of our model to test for implicit preferences along race lines. Specifically, we will test whether participants tend to award more money to applicants of one race, and less to the other, when the comparison is more opaque about race.

Participants were presented with two hypothetical application forms (constructed from real applications) side-by-side. They were asked to allocate up to a total of \$1,500 to the two applicants, with the remainder going to to "offset the deficit." We therefore interpret the decision as joint evaluation.[29]

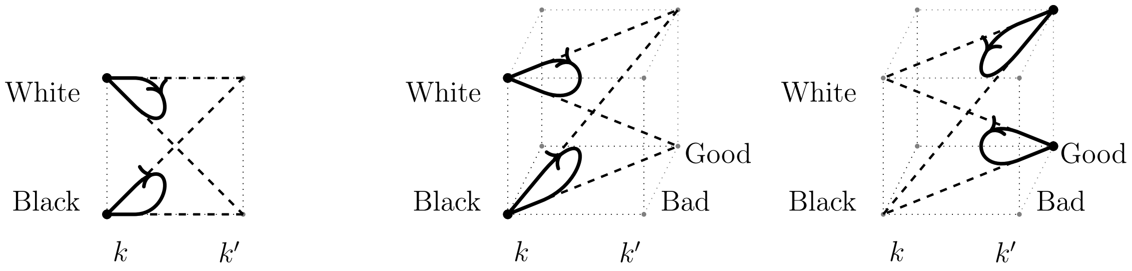
The key attribute of interest is the applicant's Race $\in$ {Black, White}, signaled by their name (Latoya and Keisha for Black applicants, Laurie and Emily for Whites). Crucially, some participants evaluate two applicants of the same Race, while others evaluate one from each Race.[30] Second, in some conditions there is also an assessment of each applicant's Work Ethic $\in$ {Good, Bad}.[31] When reported, this attribute always varies within the comparison set. Third, there are some less salient additional characteristics (e.g. the ages of the applicants' children), which are randomized independently of race and work ethic. These are not observed in the data, so we will treat them as a third "background" attribute Kids $\in \{k, k'\}$ which always differs within the comparison set, with no implicit preference attached to it.

[29] The total budget constraint introduces a slight complication, since when it binds, a participant that wants to assign a high value to one applicant is constrained to give less to the other. This could make it harder to detect implicit preferences, as it is expected to particularly constrain allocations when the comparison set contains two of the most implicitly-preferred applicants. In the data 31 percent of participants allocate the whole \$1,500 to the two applicants.

[30] A possible threat to identification, which has been pointed out in the context of other studies using names to signal race such as Bertrand and Mullainathan (2004), is that names might signal something additional to race such as social class. That could be interpreted as an additional attribute whose implicit preferences cannot be separated from those on race.

[31] The language in the experiment is "Excellent/Poor", we use "Good/Bad" for compactness. In some experimental conditions (labeled 1–3 in the paper), Race is hidden. These conditions contain no variation in opacity, so we drop them.

Figure 7 represents the data structure graphically. We observe some evaluations over bundles with two attributes (Panel (a)), and some with three (Panel (b)). Each applicant is evaluated alongside a Black comparator and a white comparator, who are otherwise identical to each other. For example, candidate (Black, Bad, k) is evaluated alongside (Black, Good, k') and (White, Good, k'). From these we can construct six convex scissors and three pairs of parallel convex scissors (see Section 3).



(a) Work ethic concealed

(b) Work ethic revealed

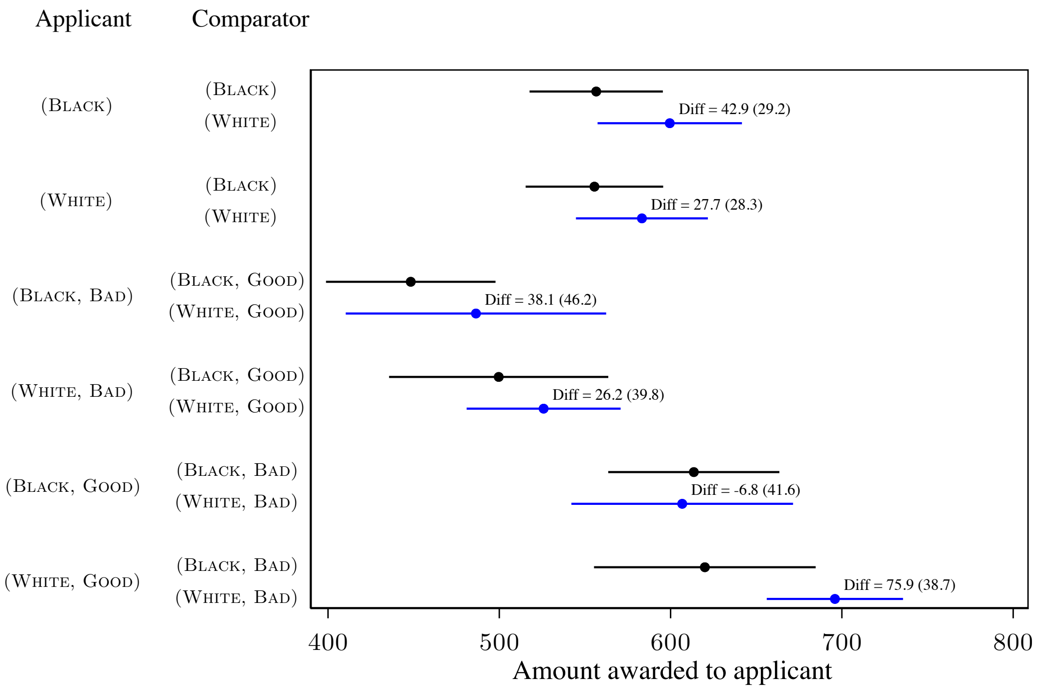
Labels k and k' indicate the background attribute Kids which varies independently of the other attributes and over which we assume there are no implicit preferences. On each diagram we draw a pair of parallel scissors which, if observed, would reveal a positive implicit preference for white applicants.

Figure 7: DeSante (2013) data structure

Our workhorse mode, Separable Implicit Preferences, does not rank opacity between these comparisons, because race switches from shared to non-shared. It can thus tell us if there is an implicit racial preference, but not its sign. We thus make use of Assumption 5. Specifically, we assume that one of the many attributes that are shared between comparisons (e.g., the fact that all applicants are female), satisfies the assumption, such that opacity is higher for shared attributes than non-shared. This tells us that we learn more about attitudes toward Black applicants relative to whites when there is one Black and one white applicant, than when both are Black or white.

Given this setup, an implicit pro-White preference will manifest as (1) Higher evaluations of Black applicants when compared to White comparators than when compared to Black comparators, and (2) Higher evaluations of white applicants when compared to white comparators than when compared to Black comparators. In other words, evaluations should always increase when the comparator switches from Black to white.

The experiment uses a between-subjects design, that is, each participant reports exactly one pair of evaluations, corresponding to one of the comparison sets in Figure 7. We therefore cannot identify implicit preferences at the individual level. Instead we will compare average evaluations across different comparison sets, and interpret these averages as revealing the preferences of a representative agent, as explained in Section 6.



Positive values of "Diff' are consistent with pro-White implicit preferences. $N = 753$ participants and 1,506 evaluations. Standard errors clustered at participant level. The average difference across conditions is 43.0 (s.e. 13.5). Joint test for no influence of any comparator, $F(6, 752) = 1.91, p = 0.0765$.

Figure 8: Reanalysis of DeSante (2013) data

We present the results in Figure 8. We group evaluations in pairs that hold fixed the target applicant, corresponding to six convex scissor tests in three pairs of parallel convex scissors. We order the data such that under implicit pro-White preferences the second (blue) evaluation should always lie to the right of the first (black).

We find positive differences in five out of six scissors, meaning that the general pattern is as predicted by implicit pro-White preferences. Only one of these individual differences is statistically significant: the evaluation of (White, Good) candidates is significantly higher when the comparator is (White, Bad) than when the comparator is (Black, Bad), but the average of the differences across conditions is $43, which is highly significantly different from zero ($p < 0.01$). An F-test of the null of no difference in any Scissor has a p-value of 0.08. Overall, we find evidence of implicit pro-White preferences in this dataset.[32]

[32] One can also use these data to test for implicit preferences over Work Ethic (when this attribute is included). Opacity about Work Ethic is higher in the comparisons where Work Ethic co-varies with Race, so implicit pro-Good preferences would manifest as higher evaluations of Good candidates in these comparisons. Figure 8 does not suggest any systematic pro-Good or pro-Bad patterns.

8 Conclusion

Our paper formalizes an assumption that is latent in a number of empirical papers: that people maintain two layers of preference for a given attribute, such that one preference (the implicit preference) becomes stronger when the comparison between outcomes becomes less direct (when it is a dilution).

By formalizing this assumption we are able to give precise instructions for inferring a person's implicit preferences from their decisions in a way that is applicable to many existing empirical datasets.

A natural formal extension would be to extend the representation theorem to other definitions of comparison, such as $\delta(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x} - \boldsymbol{z})$, which could give conditions for identification of implicit preferences in additional models such as a sophisticated signaling model.

A natural empirical extension would be to run fresh experiments designed to systematically map out the existence, strength, and consistency of implicit preferences across a range of different attributes.

References

Andreoni, J. and B. D. Bernheim (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77*(5), 1607–1636.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2020). Explicit and implicit belief-based gender discrimination: A hiring experiment. *WZP SP II 2020–306*.

Benabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies 70*(3), 489–520.

Benabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review*, 94–98.

Bertrand, M. and E. Duflo (2017). Field Experiments on Discrimination. In *Handbook of Field Experiments*, pp. 309–393. Elsevier.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review 94*(4), 991–1013.

Bohnet, I., A. van Geen, and M. Bazerman (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science 62*(5), 1225–1234.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and consumer choice. *Journal of Political Economy 121*(5), 803–843.

Border, K. C. (2020). Alternative linear inequalities.

Bushong, B., M. Rabin, and J. Schwartzstein (2020). A model of relative thinking. *The Review of Economic Studies 88*(1), 162–191.

Caplin, A. and D. Martin (2014). A testable theory of imperfect perception. *The Economic Journal 125*(582), 184–202.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics 134*(3), 1163–1224.

Chance, Z. and M. I. Norton (2009). "I Read Playboy for the Articles": Justifying and Rationalizing Questionable Preferences. In M. S. McGlone and M. L. Knapp (Eds.), *The Interplay of Truth and Deception: New Agendas in Theory and Research*, Chapter 9. Routledge.

Cherepanov, V., T. Feddersen, and A. Sandroni (2013). Rationalization. *Theoretical Economics 8*(3), 775–800.

Corno, L., E. L. Ferrara, and J. Burns (2018). Interaction, stereotypes and performance. evidence from south africa. *IFS Working Paper W19/03*.

Cubitt, R., R. McDonald, and D. Read (2018). Time matters less when outcomes differ: Unimodal vs. cross-modal comparisons in intertemporal choice. *Management Science 64*(2), 873–887.

Cunningham, T. (2013). Comparisons and choice. *Unpublished manuscript, Stockholm University*.

Cunningham, T. (2014). Hierarchical aggregation of information and decision-making.

Dana, J., R. A. Weber, and J. X. Kuang (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory 33*(1), 67–80.

DeSante, C. D. (2013). Working twice as hard to get half as far: Race, work ethic, and america's deserving poor. *American Journal of Political Science 57*(2), 342–356.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies 83*(2), 587–628.

Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *The Quarterly Journal of Economics 132*(3), 1219–1260.

Greenwald, A. G., M. R. Banaji, and B. A. Nosek (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology 108*(4), 553–561.

Greenwald, A. G. and L. H. Krieger (2006). Implicit bias: Scientific foundations. *California Law Review 94*(4), 945.

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology 74*(6), 1464–1480.

Hsee, C. K., G. F. Loewenstein, S. Blount, and M. H. Bazerman (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin 125*(5), 576–590.

Hsee, C. K. and J. Zhang (2010). General evaluability theory. *Perspectives on Psychological Science 5*(4), 343–355.

Jungnickel, D. (2005). *Graphs, Networks and Algorithms.* Springer.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kahneman, D. and S. Frederick (2005). A model of heuristic judgment. *The Cambridge handbook of thinking and reasoning*, 267–294.

Kőszegi, B. and A. Szeidl (2012). A model of focusing in economic choice. *The Quarterly Journal of Economics 128*(1), 53–104.

Manzini, P. and M. Mariotti (2007). Sequentially rationalizable choice. *American Economic Review 97*(5), 1824–1839.

Manzini, P. and M. Mariotti (2012). Choice by lexicographic semiorders. *Theoretical Economics 7*(1), 1–23.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012). Revealed attention. *American Economic Review 102*(5), 2183–2205.

Müller-Trede, J., S. Sher, and C. R. M. McKenzie (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision 2*(4), 280–305.

Norton, M. I., J. A. Vandello, and J. M. Darley (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology 87*(6), 817–831.

Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology 105*(2), 171–192.

Rand, D. G., J. D. Greene, and M. A. Nowak (2012). Spontaneous giving and calculated greed. *Nature 489*(7416), 427–430.

Regenwetter, M., J. Dana, and C. P. Davis-Stober (2011). Transitivity of preferences. *Psychological Review 118*(1), 42.

Ridout, S. (2021). Choosing for the right reasons. *Working paper*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics 50*(3), 665–690.

Snyder, M. L., R. E. Kleck, A. Strenta, and S. J. Mentzer (1979). Avoidance of the handicapped: an attributional ambiguity analysis. *Journal of personality and social psychology 37*(12), 2297.

Woodford, M. (2012). Inattentive valuation and reference-dependent choice. *Unpublished Manuscript, Columbia University*.

A Appendix

A.1 Proof of Theorem 1

We first express the rationalizability problem as a matrix of inequalities. Each inequality in the dataset can written as:

$$v(\boldsymbol{x}^j) + \sum \boldsymbol{x}^j_i \kappa_i \theta_i(\delta(\boldsymbol{x}^j, \boldsymbol{z}^j)) \ \substack{> \\ \geq}\ v(\boldsymbol{x}'^j) + \sum \boldsymbol{x}'^j_i \kappa_i \theta_i(\delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j)).$$

The unobserved functions, $v(\cdot)$ and $\theta_i(\cdot)$, can be written as vectors $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{X}|}$ and $\boldsymbol{\theta} \in \mathbb{R}^{n|\Delta|}$, with elements $v_x = v(\boldsymbol{x})$, and $\theta_{i,\delta} = \theta_i(\delta)$. We will express rationalizability of the dataset with two matrix inequalities: $\left[\begin{smallmatrix} \hat{P} & \hat{X} \end{smallmatrix}\right]\left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right] \gg 0$ representing the $\bar{m}$ strict inequalities, and $\left[\begin{smallmatrix} \bar{P} & \bar{X} \end{smallmatrix}\right]\left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right] \geq 0$ representing the $m - \bar{m}$ weak inequalities. The matrix $P = \left[\begin{smallmatrix} \hat{P} \\ \bar{P} \end{smallmatrix}\right] \in \mathbb{Z}^{m \times |\mathcal{X}|}$ holds the coefficients on $\boldsymbol{v}$, with entries:

$$P_{\underbrace{j}_{\substack{\text{row} \\ j \in 1,\ldots,m}}, \underbrace{\boldsymbol{x}}_{\substack{\text{column} \\ x \in \mathcal{X}}}} = \underbrace{\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\}}_{\text{LHS of inequality}} - \underbrace{\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}.}_{\text{RHS of inequality}}$$

The matrix $X = \left[\begin{smallmatrix} \hat{X} \\ \bar{X} \end{smallmatrix}\right] \in \mathbb{Z}^{m \times n|\mathcal{X}|}$ holds the coefficients on $\boldsymbol{\theta}$, with entries:

$$X_{\underbrace{j}_{\substack{\text{row} \\ j \in 1,\ldots,m}}, \underbrace{i\delta}_{\substack{\text{column} \\ i \in 1,\ldots,n \\ \delta \in \Delta}}} = x^j_i \kappa_i \underbrace{\mathbb{1}\{\delta = \delta(\boldsymbol{x}^j, \boldsymbol{z}^j)\}}_{\substack{=1 \text{ if LHS of inequality } j \\ \text{has comparison } \delta}} - x'^j_i \kappa_i \underbrace{\mathbb{1}\{\delta = \delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j)\}.}_{\substack{=1 \text{ if RHS of inequality } j \\ \text{has comparison } \delta}}$$

Finally matrix $Q \in \mathbb{Z}^{n|\Delta|^2 \times n|\Delta|}$ holds coefficients on $\boldsymbol{\theta}$ which encode Assumption 2. Each row represents a combination of an attribute k and two comparisons $\bar{\delta}, \bar{\delta}'$, and has non-zero entries only if $\bar{\delta} \sqsupseteq_k \bar{\delta}$:

$$Q_{\underbrace{k\bar{\delta}\bar{\delta}'}_{\substack{\text{row} \\ k \in \{1,\ldots,n\} \\ \bar{\delta},\bar{\delta}' \in \Delta}}, \underbrace{i\delta}_{\substack{\text{column} \\ i \in \{1,\ldots,n\} \\ \delta \in \Delta}}} = \mathbb{1}\{ \underbrace{(i = k)}_{\substack{\text{if column} \\ \text{corresponds to } k}} \wedge \underbrace{(\bar{\delta} \sqsupseteq_i \bar{\delta}')}_{\substack{\text{if opacity} \\ \text{dominance}}}\}(\underbrace{\mathbb{1}\{\delta = \bar{\delta}\}}_{\substack{\text{if column} \\ \text{corresponds to } \bar{\delta}}} - \underbrace{\mathbb{1}\{\delta = \bar{\delta}'\}}_{\substack{\text{if column} \\ \text{corresponds to } \bar{\delta}'}}$$

Rationalizability can then be expressed in the following condition:

Condition 1. *There exists a real-valued vector* $\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right]$ *satisfying*

$$\begin{bmatrix}\hat{P} & \hat{X}\end{bmatrix}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{bmatrix} \gg \mathbf{0} \quad \textit{(all positive)}$$
$$\begin{bmatrix}\bar{P} & \bar{X}\\ 0 & Q\end{bmatrix}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{bmatrix} \geq \mathbf{0} \quad \textit{(all non-negative).}$$

Applying Motzkin's Rational Transposition Theorem (Border (2020)) to this set of linear inequalities implies that Condition 1 will be true if and only if another condition, Condition 2, is false. Intuitively Condition 2 expresses that a non-negative combination of the rows of the matrix $\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]$ can be summed to make a row of zeroes.

Condition 2. *There exist integer-valued vectors* $\hat{\boldsymbol{p}} \in \mathbb{Z}^{\bar{m}}$*,* $\bar{\boldsymbol{p}} \in \mathbb{Z}^{m-\bar{m}}$*,* $\boldsymbol{q} \in \mathbb{Z}^{n|\Delta|}$ *(with* $\boldsymbol{p} \equiv \left[\begin{smallmatrix}\hat{p}\\ \bar{p}\end{smallmatrix}\right]$*), satisfying:*

$$\hat{\boldsymbol{p}}^T\begin{bmatrix}\hat{P} & \hat{X}\end{bmatrix} + \bar{\boldsymbol{p}}^T\begin{bmatrix}\bar{P} & \bar{X}\end{bmatrix} + \boldsymbol{q}^T\begin{bmatrix}\mathbf{0} & Q\end{bmatrix} = \begin{bmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{bmatrix}\begin{bmatrix}P & X\\ \mathbf{0} & Q\end{bmatrix} = \mathbf{0}^T,$$

$$\begin{aligned}\hat{\boldsymbol{p}} &> 0 && \textit{(all non-negative, at least one positive)}\\ \bar{\boldsymbol{p}} &\geq 0 && \textit{(all non-negative)}\\ \boldsymbol{q} &\geq 0 && \textit{(all non-negative).}\end{aligned}$$

Loosely speaking, implicit preferences $\boldsymbol{\kappa}$ can rationalize a dataset if and only if there is no combination of inequalities on $\boldsymbol{v}$ and $\boldsymbol{\theta}$ (i.e. rows in $\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]$), which exactly cancel, because that would lead to a contradiction.[33] It remains to show that Condition 2 is equivalent to the condition given in the theorem:

Condition 3. *There exists a cyclical selection* $\boldsymbol{s} \in \mathbb{N}^m$ *in which, for every* $\kappa_i = 1$*, the losses opacity-dominate the wins, and for every* $\kappa_i = -1$*, the wins opacity-dominate the losses.*

Proof that condition 3 implies condition 2. We will construct the two Motzkin vectors, $\boldsymbol{p}$ and $\boldsymbol{q}$, from the selection and the matching:

$$\forall j \in \{1,\ldots,m\}, \qquad p_j = s_j$$
$$\forall i \in \{1,\ldots,n\}, \delta, \delta' \in \Delta, \quad q_{i\delta\delta'} = M_{i,\delta,\delta'}$$

We can verify that $\hat{\boldsymbol{p}} > 0$, $\bar{\boldsymbol{p}} \geq 0$, from the definition of a cyclical selection, and $\boldsymbol{q} \geq 0$ from the definition of a matching.

[33] Condition 1 implies $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right] > \mathbf{0}$, condition 2 implies $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P & X\\ \mathbf{0} & Q\end{smallmatrix}\right]\left[\begin{smallmatrix}\boldsymbol{v}\\ \boldsymbol{\theta}\end{smallmatrix}\right] = \mathbf{0}$.

For any element of the vector $\boldsymbol{p}^T P \in \mathbb{Z}^{|\mathcal{X}|}$, we can write:

$$\sum_{j=1}^{m} p_j P_{j,\boldsymbol{x}} = \sum_{j=1}^{m} s_j P_{j,\boldsymbol{x}} = \underbrace{\sum_{j:\boldsymbol{x}^j=\boldsymbol{x}}}_{\substack{\boldsymbol{x}\text{ appears}\\ \text{on LHS of}\\ \text{inequality } j}} s_j - \underbrace{\sum_{j:\boldsymbol{x}'^j=\boldsymbol{x}}}_{\substack{\boldsymbol{x}\text{ appears}\\ \text{on RHS of}\\ \text{inequality } j}} s_j = 0.$$

Where the last step follows from the definition of a cyclical selection: each bundle $\boldsymbol{x}$ must appear equally often on the left-hand and right-hand side. Thus $\boldsymbol{p}^T P = \mathbf{0}$.

We next show that $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}X\\ Q\end{smallmatrix}\right] = \mathbf{0}$. An element of the vector $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}X\\ Q\end{smallmatrix}\right] \in \mathbb{Z}^{n|\Delta|}$, indexed by $(i\delta)$, can be expressed as:

$$\underbrace{\sum_{j=1}^{m} p_j X_{j,i\delta}}_{\text{elements of } X \text{ selected by } \boldsymbol{p}} \quad + \underbrace{\sum_{k=1}^{n}\sum_{\bar{\delta}\in\Delta}\sum_{\bar{\delta}'\in\Delta} q_{k\bar{\delta}\bar{\delta}'} Q_{k\bar{\delta}\bar{\delta}',i\delta}}_{\text{elements of Q selected by } \boldsymbol{q}} .$$

Using the definitions of X and Q we can write this as:

$$\underbrace{\sum_{j:\delta(x^j,z^j)=\delta} p_j x_i^j \kappa_i - \sum_{j:\delta(x'^j,z'^j)=\delta} p_j x_i'^j \kappa_i}_{\text{inequalities on } \boldsymbol{\theta} \text{ from selection}} + \underbrace{\sum_{\bar{\delta}':\delta\sqsupseteq_i\bar{\delta}'} q_{i,\delta,\bar{\delta}'}}_{\substack{\text{where } \delta\\ \text{dominates}}} - \underbrace{\sum_{\bar{\delta}:\bar{\delta}\sqsupseteq_i\delta} q_{i,\bar{\delta},\delta}}_{\substack{\text{where } \delta\\ \text{is dominated}}} \tag{2}$$

Given $\boldsymbol{p} = \boldsymbol{s}$ the first two terms will be equal to the score for that combination of i and δ:

$$\sum_{j:\delta(x^j,z^j)=\delta} s_j x_i^j \kappa_i - \sum_{j:\delta(x'^j,z'^j)=\delta} s_j x_i'^j \kappa_i = \kappa_i c_{i,\delta}.$$

We can then take the last two terms of (2), using $q_{i\delta\delta'} = M_{i,\delta,\delta'}$:

$$\sum_{\bar{\delta}':\delta\sqsupseteq_i\bar{\delta}'} M_{i,\delta,\bar{\delta}'} - \sum_{\bar{\delta}:\bar{\delta}\sqsupseteq_i\delta} M_{i,\bar{\delta},\delta} = \begin{cases} -c_{i,\delta} & ,\kappa_i = 1 \;\; \text{(when losses dominate wins)} \\ c_{i,\delta} & ,\kappa_i = -1 \;\; \text{(when wins dominate losses)} \end{cases}$$
$$= -\kappa_i c_{i,\delta}$$

Combined with the prior step we thus have $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P & X\\ 0 & Q\end{smallmatrix}\right] = \mathbf{0}$, establishing Condition 2.

Proof that condition 2 implies condition 3. We construct our selection vector $\boldsymbol{s}$ and a set of matrices M_i for $i = \{1, \dots, n\}$ as:

$$\forall j \in \{1, \dots, m\}, \qquad s_j = p_j$$
$$\forall i \in \{1, \dots, n\}, \delta, \delta' \in \Delta, \quad M_{i,\delta,\delta'} = \begin{cases} q_{i\delta\delta'} & , \delta \sqsupseteq_i \delta' \\ 0 & , \text{otherwise.} \end{cases}$$

We can verify that $s_j \geq 0$ and $M_{i,\delta,\delta'} \geq 0$ because $\bar{\boldsymbol{p}}, \boldsymbol{q} \geq 0$, and that $s_j > 0$ for at least one $j \leq \bar{m}$ because $\hat{\boldsymbol{p}} > 0$. To confirm that $\boldsymbol{s}$ is a cyclical selection we need to show that

$$\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} = \sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\},$$

which follows because $\boldsymbol{p}'P = \boldsymbol{0}$, and:

$$\begin{aligned} (\boldsymbol{p}'P)_{\boldsymbol{x}} &= \sum_{j=1}^{m} p_i P_{j,\boldsymbol{x}} \\ &= \sum_{j=1}^{m} p_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} - \sum_{j=1}^{m} p_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}. \end{aligned}$$

We finally verify that for each i with $\kappa_i = 1$ the losses opacity-dominate the wins, meaning the matrix M_i satisfies the conditions of Definition 3 (the same argument will show that, when $\kappa_i = -1$, the wins opacity-dominate the losses).

1. Matches obey dominance: $\forall \delta, \delta' \in \Delta$, $(M_{i,\delta,\delta'} > 0) \implies (\delta \sqsupseteq_i \delta')$. This immediately follows from our construction of M from $\boldsymbol{q}$ above.
2. All scores are accounted for, i.e. for every $\delta \in \Delta$:

$$\begin{aligned} \underbrace{\sum_{\bar{\delta}' \in \Delta} M_{i,\delta,\bar{\delta}'}}_{\delta \text{ dominating}} - \underbrace{\sum_{\bar{\delta} \in \Delta} M_{i,\bar{\delta},\delta}}_{\delta \text{ dominated}} &= \sum_{\bar{\delta}' : \delta \sqsupseteq_i \bar{\delta}'} q_{i,\delta,\bar{\delta}'} - \sum_{\bar{\delta} : \bar{\delta} \sqsupseteq_i \delta} q_{i,\bar{\delta},\delta} && \text{(from construction of } M\text{)} \\ &= (\boldsymbol{q}^T Q)_i && \text{(from definition of } Q\text{)} \\ &= -(\boldsymbol{p}^T X)_i && \text{(from condition 2)} \\ &= - \sum_{j : \delta(\boldsymbol{x}^j, \boldsymbol{z}^j) = \delta} p_j x_i^j + \sum_{j : \delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = \delta} p_j x_i'^j && \text{(from definition of } X \text{ with } \kappa_i = 1\text{)} \\ &= - \sum_{j : \delta(\boldsymbol{x}^j, \boldsymbol{z}^j) = \delta} s_j x_i^j + \sum_{j : \delta(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = \delta} s_j x_i'^j && \text{(from construction of } \boldsymbol{s}\text{)} \\ &= -c_{i,\delta} && \text{(from definition of score)} \end{aligned}$$

A.2 Proofs for Section 4 (Foundations)

Proof of Proposition 1 It is easy to see that θ_i depends only on $(\boldsymbol{x}, \boldsymbol{z})$ through $|\boldsymbol{x} - \boldsymbol{z}|$. θ_i^S is weakly increasing as the set of shared attributes grows since θ_i^S is a constant. We need to show that θ_i^N is weakly increasing as the set of non-shared attributes grows. Let $|\boldsymbol{x}' - \boldsymbol{z}'| \sqsupseteq_i |\boldsymbol{x} - \boldsymbol{z}|$. Consider the set of attributes that are shared under $|\boldsymbol{x} - \boldsymbol{z}|$ and become non-shared under $|\boldsymbol{x}' - \boldsymbol{z}'|$, i.e. $D = \{j : (j \in S^{|\boldsymbol{x}-\boldsymbol{z}|}) \wedge (j \notin S^{|\boldsymbol{x}'-\boldsymbol{z}'|})\}$. If all of them are governed by a rule $(\forall j \in D, \lambda_j \neq 0)$ then the rule-applying function is unaffected, so $\theta_i^N(|\boldsymbol{x}' - \boldsymbol{z}'|) = \theta_i^N(|\boldsymbol{x} - \boldsymbol{z}|)$. If one or more is not governed by a rule $(\exists j \in D : \lambda_j = 0)$, then $\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S^{|\boldsymbol{x}'-\boldsymbol{z}'|})\} = 0$, so $\theta_i^N(|\boldsymbol{x}' - \boldsymbol{z}'|) = |\lambda_i| \geq \theta_i^N(|\boldsymbol{x} - \boldsymbol{z}|)$.

Proof of Proposition 2 It is easy to see that θ_i depends only on $(\boldsymbol{x}, \boldsymbol{z})$ through $|\boldsymbol{x} - \boldsymbol{z}|$. We need to show that θ^S and θ^N are weakly increasing as the sets of shared and non-shared attributes grow respectively. θ^S is a constant. It is easy to see that θ^N increases as we add additional non-shared attributes.

Proof of Proposition 3 It is easy to see that θ_i depends only on $(\boldsymbol{x}, \boldsymbol{z})$ through $|\boldsymbol{x} - \boldsymbol{z}|$. It is easily seen also be seen that θ^S and θ^N are weakly increasing as we add additional shared and non-shared attributes respectively.

Proof of Proposition 4 Assign index t to the attribute that has either $\lambda_i \neq 0$ or $\pi_i \neq 1$. Our goal is to show that the functional form derived in Lemma 3 implies that of the Proposition. First, observe that $\bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|)$ can be written as;

$$\begin{aligned}
\bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|) &= \begin{cases} 1 - \frac{\sum_{j \in S}(1-\pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} & , i \in S \\ 1 - \frac{\sum_{j \notin S}(1-\pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2} & , i \notin S \end{cases} \\
&= \begin{cases} 1 - \frac{(1-\pi_i)\sigma_i^2}{\sum_{j \in S}\sigma_j^2} - \frac{\sum_{(j \in S) \wedge (j \neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} & , i \in S \\ 1 - \frac{(1-\pi_i)\sigma_i^2}{\sum_{j \notin S}\sigma_j^2} - \frac{\sum_{(j \notin S) \wedge (j \neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2} & , i \notin S \end{cases} \\
&= \begin{cases} \pi_i + (1-\pi_i)\left(1 - \frac{\sigma_i^2}{\sum_{j \in S}\sigma_j^2}\right) - \frac{\sum_{(j \in S) \wedge (j \neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} & , i \in S \\ \pi_i + (1-\pi_i)\left(1 - \frac{\sigma_i^2}{\sum_{j \notin S}\sigma_j^2}\right) - \frac{\sum_{(j \notin S) \wedge (j \neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2} & , i \notin S. \end{cases}
\end{aligned}$$

Substituting into the functional form derived in Lemma 3, we obtain:

$$u^{IA}(|\boldsymbol{x}-\boldsymbol{z}|) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i\lambda_i\pi_i + \sum_{i=1}^{n} x_i sgn(\lambda_i(1-\pi_i))\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) - B$$

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i(1-\pi_i)|\left(1-\frac{\sigma_j^2}{\sum_{j\in S}\sigma_j^2}\right) & ,i\in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i(1-\pi_i)|\left(1-\frac{\sigma_j^2}{\sum_{j\notin S}\sigma_j^2}\right) & ,i\notin S, \end{cases}$$

which uses $\lambda_i(1-\pi_i) = sgn(\lambda_i(1-\pi_i))|\lambda_i(1-\pi_i)|$, and where:

$$\begin{aligned} B &= \begin{cases} \sum_{i=1}^{n} x_i\lambda_i\frac{\sum_{(j\in S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2} & ,t\in S \\ \sum_{i=1}^{n} x_i\lambda_i\frac{\sum_{(j\notin S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} & ,t\notin S \end{cases} \\ &= \begin{cases} x_t\lambda_t\frac{\sum_{(j\in S)\wedge(j\neq t)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2} + \sum_{i\neq t} x_i\lambda_i\frac{\sum_{(j\in S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\in S}\sigma_j^2} & ,t\in S \\ x_t\lambda_t\frac{\sum_{(j\notin S)\wedge(j\neq t)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} + \sum_{i\neq t} x_i\lambda_i\frac{\sum_{(j\notin S)\wedge(j\neq i)}(1-\pi_j)\sigma_j^2}{\sum_{j\notin S}\sigma_j^2} & ,t\notin S \end{cases} \\ &= 0, \end{aligned}$$

where the final step uses the facts that $\forall j \neq t,\ \pi_j = 1$ (so the first term equals zero) and $\lambda_j = 0$ (so the second term equals zero).

To complete the proof, note that (1) θ_i depends only on $(\boldsymbol{x}, \boldsymbol{z})$ through $|\boldsymbol{x}-\boldsymbol{z}|$, and (2) θ^S and θ^N are weakly increasing as add additional shared and non-shared attributes, respectively.

Proof of Proposition 5

Ceteris paribus. No implicit preference for k means $\lambda_k = 0$.

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) = \begin{cases} \theta_i^S(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i| & ,i\in S \\ \theta_i^N(|\boldsymbol{x}-\boldsymbol{z}|) & = |\lambda_i|(1-\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j\in S)\}) & ,i\notin S. \end{cases}$$

$$\theta_i^S - \theta_i^N = |\lambda_i|\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j\in S)\}.$$

$\theta_i^S - \theta_i^N \geq 0$, so if k is shared, shared attributes have weakly higher opacity. If k is non-shared, $\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j\in S)\} = 0$ (since $\lambda_k = 0$ and $k \notin S$), so $\theta_i^S - \theta_i^N = 0$ (opacity is the same for shared and non-shared). Thus, opacity for attribute i is weakly greater when i shares status with k.

Signaling evaluation. Let $\sigma_k^2 \geq \sum_{i\neq t}\sigma_i^2$. Then we have:

$$\theta_i^S - \theta_i^N = |\lambda_i|\sigma_i^2 \left(\frac{\sum_{j\in S} \sigma_j^2 - \sum_{j\notin S} \sigma_j^2}{\sum_{j\in S} \sigma_j^2 \sum_{j\notin S} \sigma_j^2} \right),$$

which is weakly positive if $k \in S$, weakly negative if $k \notin S$. Thus, opacity for attribute i is weakly greater when i shares status with k.

Implicit associations. Let $\sigma_k^2 \geq \sum_{i\neq t} \sigma_i^2$. Assumption 5 holds trivially for all attributes j with $\lambda_j = 0$ or $\pi_j = 1$, since $\theta_j = 0$, so we only need to check if it holds for attribute t, the attribute that may have nonzero λ or non-unity π. If k coincides with t ($k = t$), 5 holds trivially. Suppose not, i.e. $t \neq k$ (and hence $\lambda_k = 0$ and $\pi_k = 1$). We have:

$$\theta_t^S - \theta_t^N = |\lambda_t(1-\pi_t)|\sigma_t^2 \left(\frac{\sum_{j\in S} \sigma_j^2 - \sum_{j\notin S} \sigma_j^2}{\sum_{j\in S} \sigma_j^2 \sum_{j\notin S} \sigma_j^2} \right),$$

which satisfies dominating opacity by the same argument as signaling evaluation.

B Web appendix

B.1 Derivations for Section 3

All of our examples can be analyzed by partitioning the set of attributes into three disjoint and collectively exhaustive "groups," A, B, C, each containing at least one attribute.[34] All attributes within a group are perfectly correlated, so we can represent them using three grouped attributes, $\boldsymbol{x} = (x_A, x_B, x_C)$.[35]

Since attributes are perfectly correlated within dimensions, they will have identical differences in a given comparison (e.g. we have $\delta_i = \delta_j, \forall i, j \in A$). Identical differences do not translate into identical opacities (the opacity function can vary in magnitude across attributes) but all opacity inequalities will be shared, so for example $\theta_i(\boldsymbol{\delta}) \geq \theta_i(\boldsymbol{\delta}') \Leftrightarrow \theta_j(\boldsymbol{\delta}) \geq \theta_j(\boldsymbol{\delta}'), \forall i, j \in A$. Therefore, rather than laboriously write out conditions and matrices for all n attributes, we can conduct all our analysis using the three grouped attributes x_A, x_B, x_C, where $x_A \kappa_A \theta_A(\boldsymbol{\delta}) := \sum_{i \in A} x_i \kappa_i \theta_i(\boldsymbol{\delta})$. Implications that we derive on a grouped attribute will imply a disjunction over all attributes within the group, namely:

$$(x_A \kappa_A = 1) \Rightarrow \left(\bigvee_{i \in A} x_i \kappa_i = 1 \right).$$

Applying the Theorem To apply the theorem to a given dataset we construct its X matrix by collecting each inequality's wins and losses, and construct its Q matrix by assembling all opacity restrictions that can be derived for the setting. We include optional rows corresponding to the Dominance of attribute k assumption (Assumption 5).

We can use a number of shortcuts to simplify the analysis. First, every cyclical selection must put the same weight on each inequality from a given cycle (otherwise $\boldsymbol{p}^T P \neq 0$), so we can without loss of generality sum the rows of X that correspond to a single cycle, collapsing it down to a single row. Second, we will eliminate columns from X and Q where all entries in X are zero (since it is always possible to set those columns to zero in a $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix}$-weighted sum of X and Q. Third, we will eliminate rows and from X and Q that have all zero entries. Fourth, we eliminate rows from Q that do not restrict any observed inequality (that is, there is no selection in X with nonzero entries in both columns that are restricted by this row of Q).

Applying the theorem requires asking the question: for a given $\boldsymbol{\kappa}$, do there exist vectors

[34] So $A \cup B \cup C = 1, \ldots, n$; $A \neq \emptyset, B \neq \emptyset, C \neq \emptyset$; and $A \cap B = A \cap C = B \cap C = \emptyset$.

[35] For example, if $A = \{1, 2, 3\}$ we might have $x_A = -1 \Leftrightarrow (x_1, x_2, x_3) = (-1, 1, -1)$ and $x_A = 1 \Leftrightarrow (x_1, x_2, x_3) = (1, -1, 1)$.

$\boldsymbol{p}, \boldsymbol{q}$ such that $[\boldsymbol{p}^T, \boldsymbol{q}^T]\left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right] = \mathbf{0}$? Our approach will be to write out the terms of the vector $[\boldsymbol{p}^T, \boldsymbol{q}^T]\left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right]$ and ask for what values of $\boldsymbol{\kappa}$ at least one term must be nonzero, meaning that $\boldsymbol{\kappa}$ is not ruled out by Theorem 1.

Figure 9 presents the matrix representation of our examples (except for the equilateral triangle, which has very different nonzero columns to the others).

Right triangle Let $A = \{i : x_i^1 \neq x_i^2\}$, $B = \{i : x_i^2 \neq x_i^3\}$, $C = \{i : x_i^1 = x_i^3\}$. So A is the set of attributes that vary in $\boldsymbol{\delta}^1$, B is the set that vary in $\boldsymbol{\delta}^2$, $A \cup B$ is the set that vary in $\boldsymbol{\delta}^3$ (the "diagonal"), and C is the set that do not vary in any $\boldsymbol{\delta}$. Orthogonality of $\boldsymbol{\delta}^1$ and $\boldsymbol{\delta}^2$ implies A, B, C are disjoint and collectively exhaustive. Collapsing the set of attributes down to these three groups, we have:

$$\boldsymbol{\delta}^1 = \left[\begin{smallmatrix} 2 \\ 0 \\ 0 \end{smallmatrix}\right], \boldsymbol{\delta}^2 = \left[\begin{smallmatrix} 0 \\ 2 \\ 0 \end{smallmatrix}\right], \text{ and } \boldsymbol{\delta}^3 = \left[\begin{smallmatrix} 2 \\ 2 \\ 0 \end{smallmatrix}\right].$$

The choice inequalities are $u(\boldsymbol{x}^1, \boldsymbol{x}^2) > u(\boldsymbol{x}^2, \boldsymbol{x}^1)$, $u(\boldsymbol{x}^2, \boldsymbol{x}^3) > u(\boldsymbol{x}^3, \boldsymbol{x}^2)$, and $u(\boldsymbol{x}^3, \boldsymbol{x}^1) > u(\boldsymbol{x}^1, \boldsymbol{x}^3)$. To construct the X matrix we need to count the wins and losses for each $i, \boldsymbol{\delta}$ pair. Beginning with the first inequality, for each attribute i, the left-hand side gives us a win if $x_i^1 = 1$ and a loss otherwise. The right-hand side gives us a loss if $x_i^2 = 1$ and a win otherwise. From the conditions defining the right triangle, we know that $x_A^1 = -x_A^2 = -x_A^3$ while $x_B^1 = x_B^2$ and $x_C^1 = x_C^2$. So, the column associated with $A, \boldsymbol{\delta}^1$ receives net wins equal to the value of $2x_A^3$, while the $B, \boldsymbol{\delta}^1$ and $C, \boldsymbol{\delta}^1$ columns equal zero. By the same argument, the entry in $B, \boldsymbol{\delta}^2$ is $-2x_B^3$, the entries in $A, \boldsymbol{\delta}^3$ and $B, \boldsymbol{\delta}^3$ are $2x_A^3$ and $2x_B^3$. All other entries are zero.

With a single cycle (Right triangle 1) there is a unique cyclical selection (up to a multiplicative constant), so we set $\boldsymbol{p} = 1$ without loss of generality, obtaining (after eliminating columns where $X = 0$):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right] = \begin{bmatrix} -2\kappa_A x_A^3 - q_1 & 2\kappa_A x_A^3 + q_1 & -2\kappa_B x_B^3 - q_2 & 2\kappa_B x_B^3 + q_2 \end{bmatrix},$$

where q_1 is the coefficient on the first row of Q and q_2 is the coefficient on the second. There exist $q_1, q_2 \geq 0$ such that this vector equals 0 if and only if:

$$\left(\kappa_A x_A^3 \leq 0\right) \wedge \left(\kappa_B x_B^3 \leq 0\right).$$

Hence, the data can be rationalized if and only if:

$$\left(\kappa_A x_A^3 = 1\right) \vee \left(\kappa_B x_B^3 = 1\right) \Leftrightarrow \bigvee_{\{i : x_i^3 \neq x_i^1\}} (\kappa_i x_i^3 = 1),$$

$$X = \begin{array}{c|ccccccccc}
 & A, \begin{bmatrix}2\\0\\0\end{bmatrix} & A, \begin{bmatrix}0\\2\\0\end{bmatrix} & A, \begin{bmatrix}2\\2\\0\end{bmatrix} & B, \begin{bmatrix}2\\0\\0\end{bmatrix} & B, \begin{bmatrix}0\\2\\0\end{bmatrix} & B, \begin{bmatrix}2\\2\\0\end{bmatrix} & C, \begin{bmatrix}2\\0\\0\end{bmatrix} & C, \begin{bmatrix}0\\2\\0\end{bmatrix} & C, \begin{bmatrix}2\\2\\0\end{bmatrix} \\
\text{Right triangle 1} & -2\kappa_A x_A^3 & 0 & 2\kappa_A x_A^3 & 0 & -2\kappa_B x_B^3 & 2\kappa_B x_B^3 & 0 & 0 & 0 \\
\text{Right triangle 2} & -2\kappa_A \bar{x}_A^3 & 0 & 2\kappa_A \bar{x}_A^3 & 0 & -2\kappa_B \bar{x}_B^3 & 2\kappa_B \bar{x}_B^3 & 0 & 0 & 0 \\
\text{RT1} + \text{RT2} & 0 & 0 & 0 & 0 & -4\kappa_B x_B^3 & 4\kappa_B x_B^3 & 0 & 0 & 0 \\
\text{Figure 8} & -4\kappa_A x_A^4 & 0 & 4\kappa_A x_A^4 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Convex scissor 1} & -\kappa_A x_A \Upsilon & 0 & \kappa_A x_A \Upsilon & -\kappa_B x_B \Upsilon & 0 & \kappa_B x_B \Upsilon & -\kappa_C x_C \Upsilon & 0 & \kappa_C x_C \Upsilon \\
\text{Convex scissor 2} & -\kappa_A \bar{x}_A \bar{\Upsilon} & 0 & \kappa_A \bar{x}_A \bar{\Upsilon} & -\kappa_B \bar{x}_B \bar{\Upsilon} & 0 & \kappa_B \bar{x}_B \bar{\Upsilon} & -\kappa_C \bar{x}_C \bar{\Upsilon} & 0 & \kappa_C \bar{x}_C \bar{\Upsilon} \\
\text{CS 1} + \text{CS 2} & \text{see notes} & 0 & \text{see notes} & \text{see notes} & 0 & \text{see notes} & \text{see notes} & 0 & \text{see notes} \\
\text{Non-convex scissor} & -\kappa_A x_A d & \kappa_A x_A d & 0 & -\kappa_B x_B d & \kappa_B x_B d & 0 & -\kappa_C x_C d & \kappa_C x_C d & 0 \\
\text{Falsifications} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

$$Q = \begin{array}{c|ccccccccc}
\begin{bmatrix}2\\2\\0\end{bmatrix} \sqsupset_A \begin{bmatrix}2\\0\\0\end{bmatrix} & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\begin{bmatrix}2\\2\\0\end{bmatrix} \sqsupset_B \begin{bmatrix}0\\2\\0\end{bmatrix} & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
\begin{bmatrix}2\\0\\0\end{bmatrix} \sqsupset_C \begin{bmatrix}2\\2\\0\end{bmatrix} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
\text{Dominance} & 0 & 0 & 0 & \Theta & 0 & -\Theta & 0 & 0 & 0
\end{array}$$

$$\begin{array}{cc|cccccc}
 & & A, \begin{bmatrix}2\\2\\0\end{bmatrix} & A, \begin{bmatrix}2\\0\\2\end{bmatrix} & B, \begin{bmatrix}2\\2\\0\end{bmatrix} & B, \begin{bmatrix}0\\2\\2\end{bmatrix} & C, \begin{bmatrix}0\\2\\2\end{bmatrix} & C, \begin{bmatrix}2\\0\\2\end{bmatrix} \\
X = & \text{Equilateral triangle} & -2\kappa_A x_A^3 & 2\kappa_A x_A^3 & 2\kappa_B x_B^3 & -2\kappa_B x_B^3 & -2\kappa_C x_C^3 & 2\kappa_C x_C^3 \\
Q = & & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Notes: (1) Υ equals the sign of the scissor's evaluation change: $\Upsilon = sgn(y^2 - y^1)$. (2) Nonzero entries for Convex scissor 1+Convex scissor 2 (CS1+CS2) equal the sum of the above two rows, and are omitted for space reasons. (3) Q matrices include only rows that restrict at least one row of X. (4) Θ captures the sign of the Dominance of attribute k assumption (Assumption 5). $\Theta = 0$ if the Dominance assumption does not apply, $\Theta = 1$ if opacity is higher for shared attributes (k is shared), $\Theta = -1$ if opacity is higher for non-shared (k is non-shared).

Figure 9: Matrix representation of examples from Section 3

where the last part follows from the definitions of $A, B, \boldsymbol{x}^1, \boldsymbol{x}^3$.

Figure 8 Let $A = \{i : x_i^1 \neq x_i^2\}$, $B = \{i : x_i^1 \neq x_i^3\}$, $C = \{i : x_i^1 = x_i^4\}$. So A is the set of attributes that vary in the odd-numbered comparisons, B is the set of additional attributes that varies in the even-numbered but not the odd-numbered comparisons (which is nonempty since the even-numbered comparisons differ on a superset of attributes), $A \cup B$ the set that vary in the even-numbered comparisons, and C the set that are shared in all comparisons. By construction, A, B, C are disjoint and collectively exhaustive. We also have:

$$\boldsymbol{\delta}^1 = \boldsymbol{\delta}^3 = \left[\begin{smallmatrix} 2 \\ 0 \\ 0 \end{smallmatrix}\right], \text{ and } \boldsymbol{\delta}^2 = \boldsymbol{\delta}^4 = \left[\begin{smallmatrix} 2 \\ 2 \\ 0 \end{smallmatrix}\right].$$

As for the right triangle, we populate the matrix X by calculating the wins and losses for each $i, \boldsymbol{\delta}$ combination. Unlike the right triangle, all comparisons are concentrated on just two $\boldsymbol{\delta}$'s. Following the same proof strategy as for the right triangle, since we have a single cycle, we set $\boldsymbol{p} = 1$ without loss of generality. We obtain (after eliminating columns where $X = 0$):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right] = \begin{bmatrix} -4\kappa_A x_A^4 - q_1 & 4\kappa_A x_A^4 + q_1 \end{bmatrix},$$

where q_1 is the coefficient on the first row of Q. By the same argument as for the right triangle, the data can be rationalized if and only if:

$$\left(\kappa_A x_A^4 = 1\right) \Leftrightarrow \bigvee_{\{i : x_i^3 \neq x_i^4\}} (\kappa_i x_i^4 = 1),$$

where the last part follows from the definitions of $A, \boldsymbol{x}^3, \boldsymbol{x}^4$.

Parallel right triangles Let:

$$\begin{aligned} A &= \{i : x_i^1 \neq x_i^2\} = \{i : \bar{x}_i^2 \neq \bar{x}_i^3\} \\ B &= \{i : x_i^2 \neq x_i^3\} = \{i : \bar{x}_i^1 \neq \bar{x}_i^2\} \\ C &= \{i : x_i^1 = x_i^3\} = \{i : \bar{x}_i^1 = \bar{x}_i^3\}. \end{aligned}$$

In words, A is the set of attributes that are not shared in $\{\boldsymbol{x}^1, \boldsymbol{x}^2\}$ and not shared in $\{\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3\}$, B is the set of attributes that are not shared in $\{\boldsymbol{x}^1, \boldsymbol{x}^2\}$ and not shared in $\{\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3\}$, and C is the set of attributes that do not vary in any comparison (note however that C might vary across the two triangles). By construction, A, B, and C are disjoint and collectively exhaustive.

As usual, we populate the second triangle's row in X by calculating the wins and losses for each $i, \bar{\boldsymbol{\delta}}$ combination, exploiting the definitions of the triangle and the sets A, B, C to express them in terms of $\bar{\boldsymbol{x}}^3$.

When the dataset consists of a pair of parallel right triangles, a cyclical selection consists of $p_1 \geq 0$ copies of the first and $p_2 \geq 0$ copies of the second, giving us (after eliminating columns where $X = 0$):

$$\begin{aligned} \begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \left[\begin{smallmatrix} X \\ Q \end{smallmatrix} \right] &= \begin{bmatrix} -W_A & W_A & -W_B & W_B \end{bmatrix} \\ W_A &= 2\kappa_A(p_1 x_A^3 + p_2 \bar{x}_A^3) + q_1 = 2\kappa_A(p_1 - p_2)x_A^3 + q_1 \\ W_2 &= 2\kappa_B(p_1 x_B^3 + p_2 \bar{x}_B^3) + q_2 = 2\kappa_B(p_1 + p_2)x_B^3 + q_2, \end{aligned}$$

where q_1 is the coefficient on the first row of Q and q_2 is the coefficient on the second. The second steps use the fact that $x_A^3 = -\bar{x}_A^3$, and $x_B^3 = \bar{x}_B^3$.[36] Thus for a given p_1, p_2, the data can be rationalized if and only if:

$$\left(\kappa_A(p_1 - p_2)x_A^3 > 0\right) \vee \left(\kappa_B(p_1 + p_2)x_B^3 > 0\right).$$

When $p_1 = p_2$ (i.e. the selection contains an equal number of each cycle), the disjunction collapses to $(\kappa_B(p_1 + p_2)x_B^3 = 1)$, so this condition must hold for the data to be rationalizable. Once this condition holds, the data can be rationalized for all p_1, p_2, so no further restrictions are obtained by considering other selections. This, plus the definition of set B, gives us the result, that a pair of parallel right triangles implies:

$$\bigvee_{i : x_i^3 \neq x_i^2} (x_i^3 \kappa_i = 1).$$

Notice additionally that the single selection containing the combined cycles ($p_1 = p_2 = 1$) completely reveals all restrictions on $\boldsymbol{\kappa}$, and is a strict refinement relative to the conjunction

[36] The definition of the parallel right triangle, condition (1) ($\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2$) implies that the pairs of bundles are identical in all attributes that are not shared within the comparison, that is, $(\boldsymbol{x}_B^2 = \bar{\boldsymbol{x}}_B^1) \wedge (\boldsymbol{x}_B^3 = \bar{\boldsymbol{x}}_B^2)$. Similarly, condition (2) ($\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3)$) implies that both comparisons are *opposite* in all all attributes that are not shared within the comparison, that is $(\boldsymbol{x}_A^1 = -\bar{\boldsymbol{x}}_A^2) \wedge (\boldsymbol{x}_A^2 = -\bar{\boldsymbol{x}}_A^3)$. Finally, the definitions of A, B, and C imply $x_A^3 = x_A^2 = -x_A^1, x_B^3 = -x_B^2 = -x_B^1, \bar{x}_A^3 = -\bar{x}_A^2 = -\bar{x}_A^1$ and $\bar{x}_B^3 = \bar{x}_B^2 = -\bar{x}_B^1$. Substitution yields $x_A^3 = -\bar{x}_A^3$, and $x_B^3 = \bar{x}_B^3$.

of the disjunctions implied by the two cycles separately, which would be:

$$\left(\bigvee_{i:x_i^3\neq x_i^2}(x_i^3\kappa_i=1)\vee\bigvee_{i:x_i^2\neq x_i^1}(x_i^3\kappa_i=1)\right)\wedge\left(\bigvee_{i:x_i^3\neq x_i^2}(x_i^3\kappa_i=1)\vee\bigvee_{i:x_i^2\neq x_i^1}(x_i^3\kappa_i-1)\right)$$

$$\bigvee_{i:x_i^3\neq x_i^2}(x_i^3\kappa_i=1)\vee\left(\bigvee_{i:x_i^2\neq x_i^1}(x_i^3\kappa_i=1)\wedge\bigvee_{i:x_i^2\neq x_i^1}(x_i^3\kappa_i-1)\right).$$

Convex scissor without and with Dominance of attribute k. Let $A=\{i:x_i\neq z_i^1\}$, $B=\{i:z_i^1\neq z_i^2\}$, $C=\{i:x_i=z_i^2\}$. So A is the set of attributes that vary in the first comparison, B is the set of additional attributes that varies in the second comparison but not the first (which is nonempty since the second comparison differs on a superset of attributes), $A\cup B$ the full set that vary in the second comparison, and C the set that do not vary within either comparison. By construction, A,B,C are disjoint and collectively exhaustive.

We construct the scissor's row in the X matrix by counting losses and wins in the scissor's single inequality. If $y^2>y^1$ the left-hand side corresponds to $\boldsymbol{\delta}^2=\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right]$, giving us a win in column i, $\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right]$ if $\kappa_i x_i=1$ and a loss otherwise. The right-hand side corresponds to $\boldsymbol{\delta}^1=\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right]$, giving us a loss in column i, $\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right]$ if $\kappa_i x_i=1$ and a win otherwise. If $y^2<y^1$ the signs of these implications are reversed. Thus, defining $\Upsilon=sgn(y^2-y^1)$, we have $\kappa_i x_i\Upsilon$ net wins in the columns associated with $\boldsymbol{\delta}^2$, and $-\kappa_i x_i\Upsilon$ net wins in the columns associated with $\boldsymbol{\delta}^1$.

Thus, we obtain (after eliminating columns where $X=0$):

$$\begin{aligned}\begin{bmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{bmatrix}\left[\begin{smallmatrix}X\\Q\end{smallmatrix}\right]&=\begin{bmatrix}-W_A & W_A & -W_B & W_B & -W_C & W_C\end{bmatrix}\\ W_A&=\kappa_A x_A\Upsilon+q_1\\ W_B&=\kappa_B x_B\Upsilon-\Theta q_4\\ W_C&=\kappa_C x_C\Upsilon-q_3.\end{aligned}$$

where q_1 and q_3 are the coefficients on the first, and third rows of Q, which correspond to the main Opacity assumption (Assumption 2), while q_4 is the coefficient on Q's fourth row which captures the Dominance of attribute k assumption (Assumption 5). Θ embeds the assumption: $\Theta=0$ if the assumption does not apply, $\Theta=1$ if opacity is higher for shared attributes (k is shared), $\Theta=-1$ if opacity is higher for non-shared attributes (k is non-shared).

When Assumption 5 does not hold, $\Theta=0$, so the data can be rationalized if and only if:

$$(\kappa_A x_A\Upsilon=1)\vee(\kappa_B x_B\Upsilon\neq 0)\vee(\kappa_C x_C\Upsilon<0).$$

Expanding this expression using the definitions of $A, B, C,$ and Υ gives the result.

When Assumption 5 holds, $\Theta \neq 0$, so the data can be rationalized if and only if:

$$(\kappa_A x_A \Upsilon = 1) \vee (\kappa_B x_B \Upsilon = -\Theta) \vee (\kappa_C x_C \Upsilon < 0)\,.$$

Expanding this expression using the definitions of A, B, C, Υ and Θ gives the result.

Parallel convex scissors without and with Dominance of attribute k. Let:

$$\begin{aligned} A &= \{i : x_i \neq z_i^1\} = \{i : \bar{x}_i \neq \bar{z}_i^1\} \\ B &= \{i : z_i^1 \neq z_i^2\} = \{i : \bar{z}_i^1 \neq \bar{z}_i^2\} \\ C &= \{i : x_i = z_i^2\} = \{i : \bar{x}_i = \bar{z}_i^2\}. \end{aligned}$$

So A is the set of attributes that vary in each scissor's first comparison, B is the set of additional attributes that varies in the second comparisons but not the first (which is nonempty since the second comparisons differ on a superset of attributes), $A \cup B$ the full set that vary in the second comparisons, and C the set that do not vary in any comparisons (note however that C might vary across the two triangles). By construction, A, B, C are disjoint and collectively exhaustive. Since the values of $\boldsymbol{x}, \bar{\boldsymbol{x}}, sgn(y^1 - y^2)$ and $sgn(\bar{y}^1 - \bar{y}^2)$ are unrestricted, there are many possible combinations of parallel convex scissor.

When the dataset consists of a pair of parallel convex scissors, a cyclical selection consists of $p_1 \geq 0$ copies of the first and $p_2 \geq 0$ copies of the second, giving us (after eliminating columns where $X = 0$):

$$\begin{aligned} \begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \left[\begin{smallmatrix} X \\ Q \end{smallmatrix}\right] &= \begin{bmatrix} -W_A & W_A & -W_B & W_B & -W_C & W_C \end{bmatrix} \\ W_A &= \kappa_A (p_1 x_A \Upsilon + p_2 \bar{x}_A \bar{\Upsilon}) + q_1 \\ W_2 &= \kappa_B (p_1 x_B \Upsilon + p_2 \bar{x}_B \bar{\Upsilon}) - \Theta q_4 \\ W_3 &= \kappa_C (p_1 x_C \Upsilon + p_2 \bar{x}_C \bar{\Upsilon}) - q_3, \end{aligned}$$

where q_1 and q_3 are the coefficients on the first and third rows of Q, which correspond to the main Opacity assumption (Assumption 2), while q_4 is the coefficient on Q's fourth row which captures the Assumption 5. $\Upsilon = sgn(y^2 - y^1)$ and $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^1)$ capture the direction in which each evaluation changes when the comparator changes. Θ embeds Assumption k as above.

By a similar argument to the parallel right triangles, the strongest restrictions on $\boldsymbol{\kappa}$ will be obtained when $p_1 = p_2$. This maximizes the number of terms in the disjunction that become zero and drop out, and by so doing, reveals the set of restrictions that must hold in every selection. In other words, we can without loss of generality consider only the selection consisting of exactly one copy of each scissor ($p_1 = p_2 = 1$).

When Assumption 5 does not hold, $\Theta = 0$, so the data can be rationalized if and only if:

$$\big(\kappa_A(x_A\Upsilon + \bar{x}_A\bar{\Upsilon}) = 1\big) \vee \big(\kappa_B(x_B\Upsilon + \bar{x}_B\bar{\Upsilon}) \neq 0\big) \vee \big(\kappa_C(x_C\Upsilon + \bar{x}_C\bar{\Upsilon}) < 0\big)\,.$$

When Assumption 5 holds, $\Theta \neq 0$, so the data can be rationalized if and only if:

$$\big(\kappa_A(x_A\Upsilon + \bar{x}_A\bar{\Upsilon}) = 1\big) \vee \big(\kappa_B(x_B\Upsilon + \bar{x}_B\bar{\Upsilon}) = -\Theta\big) \vee \big(\kappa_C(x_C\Upsilon + \bar{x}_C\bar{\Upsilon}) < 0\big)\,.$$

Expanding the expressions using the definitions of $A, B, C, \Upsilon, \bar{\Upsilon}$ and Θ gives the results.

Note that in each case, the term corresponding to $i \in \{A, B, C\}$ is eliminated if:

$$x_i\Upsilon = -\bar{x}_i\bar{\Upsilon},$$

that is, if either (i) the second scissor has an opposite realization of x_i but evaluation moves in the same direction, or (ii) the second scissor has an identical realization of x_i, but evaluation moves in the opposite direction. Note that Assumption 5 is not required to eliminate the term for group B. Intuitively, just like for our other refinements (figure 8, parallel right triangles), these cases allow us to eliminate where preferences reverse in contradictory directions, because these attributes alone cannot explain the observed behavior. Unique identification of an implicit preference is is possible if we can eliminate all but one attribute in this way.

B.2 Proofs of Lemmas used in Section 4

In proving these lemmas we make use of an additional one that we call "Sums and Differences."

Lemma 4. *Suppose we observe two linear combinations of n independent Normal variables ("weights"), with $+1$ or -1 coefficients ("attributes"):*

$$\underbrace{\begin{bmatrix} \bar{y}^x \\ \bar{y}^z \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} x_1 & \dots & x_n \\ z_1 & \dots & z_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}}_{\boldsymbol{w}}$$

$$x_i, z_i \in \{-1, 1\}$$
$$\boldsymbol{w} = N(0, diag(\sigma_1^2, \dots, \sigma_n^2)),$$

The Bayesian posterior for unobserved weight w_i, given observed $\boldsymbol{y}$ will be:

$$E[w_i|\boldsymbol{y}] = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\frac{\bar{y}^x+\bar{y}^z}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\frac{\bar{y}^x-\bar{y}^z}{2} & , i \notin S \end{cases}.$$

In words: we can divide attributes into shared and non-shared. Our estimates of the weights on each of the shared attributes will depend only on the sum of the evaluations ($\bar{y}^x + \bar{y}^z$), and our estimates of the weights on each of the non-shared attributes will depend only on the difference between the evaluations ($\bar{y}^x - \bar{y}^z$).

Proof of Lemma 4 First we assume there exists at least one shared and one non-shared attribute (in other words, $\boldsymbol{x} \neq \boldsymbol{z}$ and $\boldsymbol{x} \neq -\boldsymbol{z}$). Given two multivariate Normals, $\boldsymbol{a}$ and $\boldsymbol{b}$, with covariance matrix: $Var\left[\begin{smallmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{smallmatrix}\right] = \left[\begin{smallmatrix}\Sigma_a & \Sigma_{a,b}\\ \Sigma'_{a,b} & \Sigma_b\end{smallmatrix}\right]$ we can write the conditional expectation: $E[\boldsymbol{a}|\boldsymbol{b}] = E[\boldsymbol{a}] + \Sigma_{a,b}\Sigma_b^{-1}(\boldsymbol{b} - E[\boldsymbol{b}])$. In our case this implies:

$$E[\boldsymbol{w}|\boldsymbol{y}] = \Sigma_{w,y}\Sigma_y^{-1}\boldsymbol{y} \tag{3}$$

with components as follows:

$$\begin{aligned}
\Sigma_y &= X\Sigma_w X' \\
&= \begin{bmatrix} \sum_i x_i^2\sigma_i^2 & \sum_i x_i z_i \sigma_i^2 \\ \sum_i x_i z_i\sigma_i^2 & \sum_i z_i^2\sigma_i^2 \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 & \sum_{i\in S}\sigma_i^2 - \sum_{i\notin S}\sigma_i^2 \\ \sum_{i\in S}\sigma_i^2 - \sum_{i\notin S}\sigma_i^2 & \sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 \end{bmatrix} \\
\Sigma_y^{-1} &= \frac{1}{4\sum_{i\in S}\sigma_i^2\sum_{i\notin S}\sigma_i^2} \begin{bmatrix} \sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 & -\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 \\ -\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 & \sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2 \end{bmatrix} \\
\Sigma_y^{-1}\boldsymbol{y} &= \frac{1}{4\sum_{i\in S}\sigma_i^2\sum_{i\notin S}\sigma_i^2} \begin{bmatrix} \left(\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2\right)\bar{y}^x + \left(-\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2\right)\bar{y}^z \\ \left(-\sum_{i\in S}\sigma_i^2 + \sum_{i\notin S}\sigma_i^2\right)\bar{y}^x + \left(\sum_{i\in S} + \sigma_i^2\sum_{i\notin S}\sigma_i^2\right)\bar{y}^z \end{bmatrix} \\
&= \frac{1}{4}\begin{bmatrix} \frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2} + \frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2} \\ \frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2} - \frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2} \end{bmatrix} \\
\Sigma_{w,y} &= \Sigma_w X' = \begin{bmatrix} x_1\sigma_1^2 & z_1\sigma_1^2 \\ \vdots & \vdots \\ x_n\sigma_n^2 & z_n\sigma_n^2 \end{bmatrix}
\end{aligned}$$

Thus, given (3), we obtain:

$$
\begin{aligned}
E[w_i|\boldsymbol{y}] &= \frac{1}{4}\left(\frac{\bar{y}^x+\bar{y}^z}{\sum_{j\in S}\sigma_j^2}+\frac{\bar{y}^x-\bar{y}^z}{\sum_{j\notin S}\sigma_j^2}\right)x_i\sigma_i^2+\frac{1}{4}\left(\frac{\bar{y}^x+\bar{y}^z}{\sum_{i\in S}\sigma_i^2}-\frac{\bar{y}^x-\bar{y}^z}{\sum_{i\notin S}\sigma_i^2}\right)z_i\sigma_i^2 \\
&= \begin{cases} x_i\frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\frac{\bar{y}^x+\bar{y}^z}{2} & , i\in S \\ x_i\frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\frac{\bar{y}^x-\bar{y}^z}{2} & , i\notin S \end{cases}
\end{aligned}
$$

Where the last step uses $x_i+z_i=2x_i\mathbf{1}\{i\in S\}$ and $x_i-z_i=2x_i\mathbf{1}\{i\notin S\}$.

Next we consider the two special cases we initially ruled out, $\boldsymbol{x}=\boldsymbol{z}$ and $\boldsymbol{x}=-\boldsymbol{z}$. We cannot use equation (3) because X does not have full rank so Σ_y is not invertible. However we can show that the same formula for $E[w_i|\boldsymbol{y}]$ applies to these cases. If all attributes are shared ($\boldsymbol{x}=\boldsymbol{z}$) then it becomes a Normal updating problem with a single observable, $\bar{y}^x=\bar{y}^z$, and each weight is updated in proportion to its share of the total variance:

$$
E[w_i|\boldsymbol{y}]=x_i\frac{\sigma_i^2}{\sum_{j=1}^n\sigma_j^2}\bar{y}^x=x_i\frac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\frac{\bar{y}^x+\bar{y}^z}{2},
$$

as in the statement of the Lemma. If instead all attributes are non-shared ($\boldsymbol{x}=-\boldsymbol{z}$) then $\bar{y}^x=-\bar{y}^z$ and we have:

$$
E[w_i|\boldsymbol{y}]=x_i\frac{\sigma_i^2}{\sum_{j=1}^n\sigma_j^2}\bar{y}^x=x_i\frac{\sigma_i^2}{\sum_{j\notin S}\sigma_j^2}\frac{\bar{y}^x-\bar{y}^z}{2},
$$

once again corresponding to the statement of the Lemma.

Proof of Lemma 1 The expectation of a Normally-distributed variable, b, conditioning on another Normal variable, a, exceeding some threshold $\bar{a}$ can be written as:

$$
E[b|a>\bar{a}]=\mu_b+\frac{\text{Cov}(a,b)}{\sqrt{Var(a)}}\frac{\phi(\frac{\bar{a}-\mu_a}{\sqrt{Var(a)}})}{1-\Phi(\frac{\bar{a}-\mu_a}{\sqrt{Var(a)}})}.
$$

In our model each w_i is Normally distributed, implying the difference in intrinsic utility between $\boldsymbol{x}$ and $\boldsymbol{z}$ will also be Normal, and so given $\boldsymbol{x}$ is chosen over $\boldsymbol{z}$ we have:

$$\begin{aligned}
E\left[w_i \middle| \sum_{j=1}^n w_j(x_j - z_j) > 0\right] &= E[w_i] + \frac{Cov(w_i, \sum_{j=1}^n w_j(x_j - z_j))}{\sqrt{Var(\sum_{j=1}^n w_j(x_j - z_j))}} \frac{\phi(0)}{1 - \Phi(0)} \\
&= \frac{(x_i - z_i)\sigma_i^2}{\sqrt{\sum_{j=1}^n (x_j - z_j)^2 \sigma_j^2}} \frac{\phi(0)}{1 - \Phi(0)} \\
&= \mathbf{1}\{i \notin S\} \frac{x_i \sigma_i^2}{\sqrt{\sum_{j \notin S} \sigma_j^2}} \frac{\phi(0)}{1 - \Phi(0)},
\end{aligned}$$

since $(x_i - z_i) = 2x_i \times \mathbf{1}\{i \notin S\}$ and $(x_i - z_i)^2 = 4 \times \mathbf{1}\{i \notin S\}$.

Proof of Lemma 2 We first define the *residual* evaluations $\bar{y}^x, \bar{y}^z$, after subtracting components which are common knowledge. For a bundle $\boldsymbol{x}$ define $\bar{y}^x$ as:

$$\bar{y}^x = y^x - g(\boldsymbol{x}) - \sum_{i=1}^n x_i k_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\} \sum_{j \in S} \sigma_j^2 + \mathbf{1}\{i \notin S\} \sum_{j \notin S} \sigma_j^2} = \sum_{i=1}^n x_i w_i$$

Next, we show that player 1's strategy $y^x = u^{SE}(\boldsymbol{x}, \boldsymbol{z})$, $y^z = u^{SE}(\boldsymbol{z}, \boldsymbol{x})$ is optimal assuming that player 2's strategy is:

$$\hat{w}_i(y^x, y^z) = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\bar{y}^x + \bar{y}^z}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\bar{y}^x - \bar{y}^z}{2} & , i \notin S \end{cases}$$

Taking first-order conditions of U^1 with respect to y^x and y^z gives us optimal values of y^x and y^z:

$$\begin{aligned}
y^x(\boldsymbol{x}, \boldsymbol{z}) &= g(\boldsymbol{x}) + \sum_{i=1}^n x_i w_i + \sum_{i=1}^n k_i \frac{\partial \hat{w}_i(y^x, y^z)}{\partial y^x} \\
&= g(\boldsymbol{x}) + \sum_{i=1}^n x_i w_i + \sum_{i=1}^n x_i k_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\} \sum_{j \in S} \sigma_j^2 + \mathbf{1}\{i \notin S\} \sum_{j \notin S} \sigma_j^2} \\
y^z(\boldsymbol{z}, \boldsymbol{x}) &= g(\boldsymbol{z}) + \sum_{i=1}^n z_i w_i + \sum_{i=1}^n z_i k_i \frac{\sigma_i^2}{\mathbf{1}\{i \in S\} \sum_{j \in S} \sigma_j^2 + \mathbf{1}\{i \notin S\} \sum_{j \notin S} \sigma_j^2}.
\end{aligned}$$

Hence $y^x(\boldsymbol{x}, \boldsymbol{z}) = u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ and $y^z(\boldsymbol{z}, \boldsymbol{x}) = u^{SE}(\boldsymbol{z}, \boldsymbol{x})$ as stated in the proposition.

Next we show that player 2's strategy is optimal given player 1's. We make use of Lemma 4 which concerns updating from a pair of binary sums of independent Gaussian variables:

the signals can be divided into the sum $(y^x + y^z)$ and difference $(y^x - y^z)$, which will be independent sufficient statistics.

Using the Lemma we can derive Player 2's optimal strategy. Taking first order conditions of U^2, we obtain the desired result:

$$\hat{w}_i(y^x, y^z) = E[w_i | y^x, y^z] = \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\bar{y}^x + \bar{y}^z}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\bar{y}^x - \bar{y}^z}{2} & , i \notin S. \end{cases}$$

Proof of Lemma 3 Given agent 1's prior on $\boldsymbol{\pi}$, we have:

$$\hat{f}(\boldsymbol{x}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i k_i E[\pi_i] = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i k_i.$$

Next, define the residual value $\bar{\hat{f}}(\boldsymbol{x})$ as follows:

$$\bar{\hat{f}}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) = \sum_{i=1}^{n} x_i k_i.$$

The second agent's posteriors for each k_i can be derived using Lemma 4:

$$\begin{aligned} E[k_i | \hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})] &= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\bar{\hat{f}}(\boldsymbol{x}) + \bar{\hat{f}}(\boldsymbol{z})}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\bar{\hat{f}}(\boldsymbol{x}) - \bar{\hat{f}}(\boldsymbol{z})}{2} & , i \notin S \end{cases} \\ &= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\sum_{i=1}^{n} (x_i + z_i) k_i}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\sum_{i=1}^{n} (x_i - z_i) k_i}{2} & , i \notin S \end{cases} \\ &= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} x_j k_j & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \sum_{j \notin S} x_j k_j & , i \notin S \end{cases} \end{aligned}$$

where the final step uses $x_i - z_i = 2x_i \mathbf{1}\{i \notin S\}$ and $x_i + z_i = 2x_i \mathbf{1}\{i \in S\}$.

The second agent thus updates on the weights on non-shared attributes from the difference between the first agent's reports for $\boldsymbol{x}$ and $\boldsymbol{z}$, and updates on the shared attributes from the sum of the reports.

The second agent's overall evaluation of bundle $\boldsymbol{x}$ will thus be equal to:

$$\begin{aligned} E[f(\boldsymbol{x})|\boldsymbol{\pi}, \hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})] &= g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \pi_i E[k_i|\hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})] \\ &= g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i^2 \pi_i \sigma_i^2 \left(\frac{\mathbf{1}\{i \in S\}}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} x_j k_j + \frac{\mathbf{1}\{i \notin S\}}{\sum_{j \notin S} \sigma_j^2} \sum_{j \notin S} x_j k_j \right) \\ &= g(\boldsymbol{x}) + \frac{\sum_{i \in S} \pi_i \sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} x_j k_j + \frac{\sum_{i \notin S} \pi_i \sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \sum_{j \notin S} x_j k_j \\ &= g(\boldsymbol{x}) + \sum_{i \in S} x_i k_i \frac{\sum_{j \in S} \pi_j \sigma_j^2}{\sum_{j \in S} \sigma_j^2} + \sum_{i \in S} x_i k_i \frac{\sum_{j \notin S} \pi_j \sigma_j^2}{\sum_{j \notin S} \sigma_j^2} = u^{IK}(\boldsymbol{x}, \boldsymbol{z}), \end{aligned}$$

where we use the convention $\mathbf{1}\{i \in S\} / \sum_{j \in S} \sigma_j^2 = 0$ if there are no shared attributes, and equivalently for the non-shared (this saves us from explicitly writing out the special cases of all shared or all non-shared attributes). The third step uses $x_i^2 = 1$ and the final step uses a switch of index labels.

B.3 Data appendix for analysis of Exley (2016)

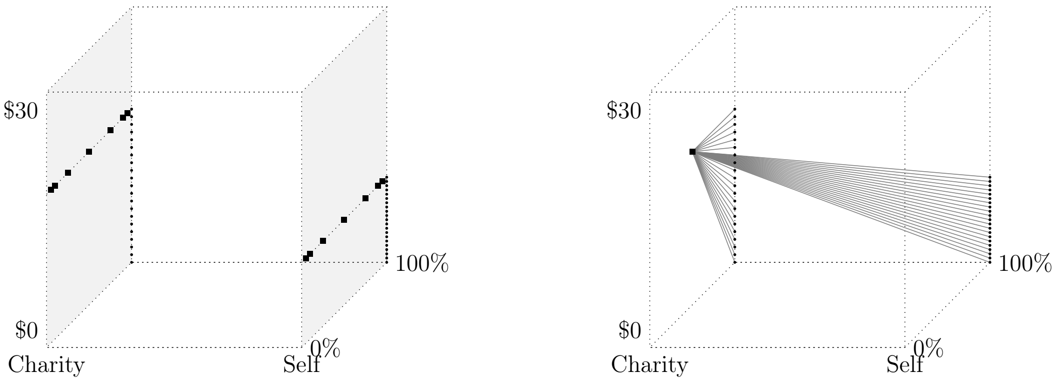
Exley (2016)'s experiment proceeds in three steps:

1. **Normalization choice.** For each participant she elicits using a choice list the smallest sure payment $\$X \in \{0, 2, \ldots, 30\}$ to charity that is chosen over \$10 for self.[37]
2. Using X, she constructs a sequence of participant-specific simple lotteries. These pay out with probability $P \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. Self lotteries, denoted by P^S, pay \$10 to self. Charity lotteries, denoted by P^C, pay $\$X$ to charity.
3. She elicits, using choice lists, preferences between each lottery and 21 different sure payoffs to self or to charity. We index these by $t = 0, ..., 20$. The sure payments are $Y_t^S = (0, 0.50, ..., 10)$ for self lotteries and $Y_t^C = (0, X/20, ..., X)$ for charity lotteries.

Thus, a bundle in this experiment is characterized by three basic attributes: a Recipient (Self or Charity), a Prize, and a Probability. Figure 10a shows graphically the full set of bundles that appear in the choice lists. Figure 10b shows every choice set a participant faces for one charity lottery.

Figure 10b shows that we do not observe all possible choices over the bundles marked in Figure 10a. Specifically, we do not observe:

[37] She conducts a second experiment where the non-Self beneficiary is another participant rather than Charity. We pool the data and simply refer to Self and Charity.



(a) All lotteries and safe payments

(b) All choice sets for one charity lottery ($P = 0.5$)

Figure 10: Exley (2016) data structure

Square markers correspond to lotteries paying charity (left face, prize $\$X$) and self (right face, prize \$10) used in the experiment. Circular markers correspond to sure payoffs on the choice lists. Any possible combination of (Recipient, Prize, Probability) lies in the shaded gray region.

- Choices between sure payoffs to Self and Charity, except for the normalization choices.
- Choices between Self and Charity lotteries.

B.3.1 Domain-specific assumptions

Exley provides some assumptions on preferences to guide her analysis and formulate her hypotheses. We will make use of the same assumptions in two ways.

1. To support our representation of the choice data in a space of two binary attributes: Social $\in$ {Selfish, Generous} and Risk $\in$ {Safe, Risky} (Section B.3.2).
2. To impute some choices that are not observed in the data (Section B.3.3).

Exley's first assumption is *linearity in payoffs*, meaning that preferences over sure payoffs are preserved under linear rescaling. So, if the participant is indifferent between $\$y$ for Charity and $\$y'$ for Self, she is also indifferent between $\$yr$ for Charity and $\$y'r$ for Self, for $r \geq 0$.

This assumption plays an important role in Exley's analysis. Her tests involve comparing certainty equivalents of Self and Charity lotteries, measured in terms of sure payments to Self and Charity. To establish inconsistencies in these certainty equivalents, she needs to be able to compare these certainty equivalents. For example, to say that the participant values a given lottery more in Self dollars than in Charity dollars, she needs to be able to rank certainty equivalents measured in these units.

The second assumption which we will refer to as *separability in probabilities* is that preferences over bundles are preserved linear rescaling of probabilities. So, if the participant is indifferent between \$$y$ for Charity and \$$y'$ for Self, she is also indifferent between \$$y$ for Charity with probability p and \$$y'$ for Self with probability p, for $p \in [0, 1]$.

These two assumptions together constitute Exley's null hypothesis, *standard risk preferences.*

Define a binary variable $c \in \{0, 1\}$ equal to one if the Recipient is Charity, and denote the Prize by y and Probability by p. The assumptions imply a utility function of the following form (normalizing $u(0) = 0$, $\pi(0) = 0$, and $\pi(1) = 1$):

$$u(c, y, p) = \pi(p)u\left(\frac{y}{1 + \lambda c}\right)$$

Linearity in payoffs is captured by λ. The participant is indifferent between a prize of y to Self and $(1 + \lambda)y$ to Charity. Separability in probabilities is captured via the probability weighting function $\pi(p)$. Preferences between two same-probability lotteries do not depend on p. Note that since all lotteries have exactly one non-zero prize, the assumption does not require *linearity* in probabilities.

To these assumptions, we add Constant Relative Risk Aversion (CRRA): $u(y) = y^\alpha$. The utility function becomes:

$$u(c, y, p) = \pi(p)\left(\frac{y}{1 + \lambda c}\right)^\alpha \tag{4}$$

We show in section B.3.4 below this utility function implies preferences have a linear separable representation in the transformed attributes Social and Risk, and thus are compatible with all of our foundations.

B.3.2 Constructing a binary attribute space

We describe here how using these assumptions we can describe the environment in terms of two binary attributes: Social $\in$ {Selfish,Generous} and Risk $\in$ {Safe, Risky}. Our approach amounts to selecting four points in the shaded region in Figure 10a, which can be described by the binary attributes Social and Risk, and over which we observe preferences. Figure 11 panels (a) and (b) provide an example.

First, note that all three attributes are "ranked": all else equal, we would expect the participant to prefer Self over Charity, and prefer larger Prizes or Probabilities to smaller. Therefore, choice sets that vary on only one of these dimensions at a time cannot satisfy Ambivalence: an observer would have strong priors on what the participant would choose.

Second, note that Prize and Probability are multivalued, and so cannot immediately fit into a binary attribute representation.

Our first step is to analyze preferences within a set of choice lists defined by the lottery probability P. We cannot make comparisons across values of P, because we do not observe such choices and because we would not expect ambivalence to be satisfied. Thus, we will construct a separate binary attribute space for each value of P. Such a space contains two probability values: lotteries with probability P, and sure payoffs with probability 1.

Second, we divide up the Prize dimension, so that Self prizes are different to Charity prizes, and sure prizes are different to risky ones, in such a way that an observer would expect the participant to be close to indifferent across any two bundles. We are guided by the assumptions described in section B.3.1. In essence we assume that (4) represents the observer's prior over the participant's preferences.

We want to identify and focus our analysis on choice sets where the participant is expected to be close to indifferent. Consider the self lottery $(0, 10, P)$ that pays \$10 to Self with probability P. Equation (4) implies the following utilities are equal:

$$u\underbrace{(0,10,P)}_{\text{Self lottery}} = u\underbrace{(1,(1+\lambda)10,P)}_{\text{Charity lottery}} = u\underbrace{\left(0,\pi(P)^{\frac{1}{\alpha}}10,1\right)}_{\text{Self sure payoff}} = u\underbrace{\left(1,(1+\lambda)\pi(P)^{\frac{1}{\alpha}}10,1\right)}_{\text{Charity sure payoff}} \tag{5}$$

The scalar $(1+\lambda)$ captures the trade-off between Charity and Self payoffs, while $\pi(P)^{\frac{1}{\alpha}}$ captures the trade-off between risky and sure payoffs.

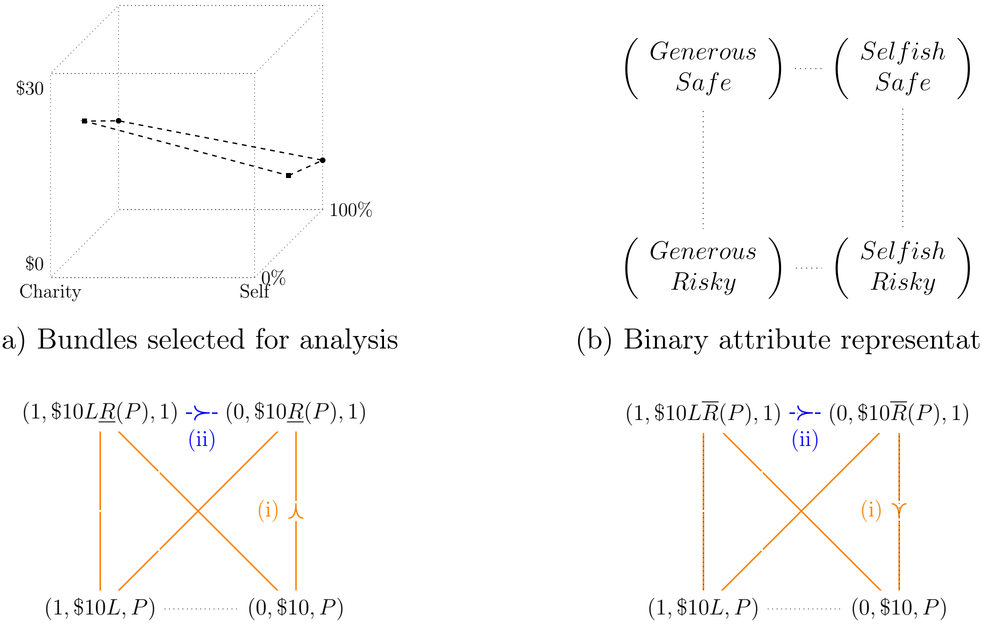
Our approach will be to focus on choices defined by two participant-specific scaling parameters, L and $R(P)$, such that Charity prizes are an L-multiple of self prizes, and sure prizes are an $R(P)$-multiple of risky prizes (with probability P). So, we will consider the Self lottery paying \$10 with probability P, the Charity lottery paying \$$10L$ with probability P, the Self sure payment of \$$10R(P)$, and the Charity sure payment of \$$10LR(P)$. Ambivalence holds if $L \approx 1+\lambda$ and $R(P) \approx \pi(P)^{\frac{1}{\alpha}}$.

We calibrate L and $R(P)$ using the participant's actual choices. L is fixed using their initial "normalization choice" between a sure payoff to self and to charity, following Exley. $R(P)$ is calibrated for each probability P using the participant's direct choices between sure payoffs to self and self lotteries. This seems a reasonable approach to picking calibration parameters where we expect the participant would be close to indifferent.

L is set using the initial normalization choice in the experiment: $L = {}^{X}\!/_{10}$. We do this because all self and charity lottery prizes differ by this ratio, so we do not observe any choices that could inform us about lottery preferences with a different prize ratio. Recall that X is the smallest payment to charity that was chosen over \$10 to self, from which we infer ${}^{X}\!/_{10} > 1+\lambda > {}^{X-2}\!/_{10}$. Linearity in payoffs and standard risk preferences therefore imply the participant can be expected to be close to indifferent between self and charity lotteries or self and charity sure payoffs differing by a ratio of L.

We consider two possible values for $R(P)$ using the participant's own choices between the self lottery and self sure payoffs. The first is based on the largest self sure payment that the participant rejected, which we denote by $\underline{Y}(P^S)$ and set $\underline{R}(P) = {}^{\underline{Y}(P^S)}/_{10}$. The second is based on the smallest self sure payoff that they accepted, which we denote by $\overline{Y}(P^S)$ (which is equal to $\overline{Y}(P^S) = \underline{Y}(P^S) + \$0.50$). This gives us $\overline{R}(P) = {}^{\overline{Y}(P^S)}/_{10}$.

Since $\underline{R}(P)$ and $\overline{R}(P)$ are close to one another, we assume that the choices based on these parameters are informative about the same binary attribute space, depicted in Figure 11b. Choice sets calibrated based on $\underline{R}(P)$ allow us to observe cycles in which (Selfish, Risky) is chosen over (Selfish, Safe) (Figure 11c). Choice sets calibrated based on $\underline{R}(P)$ allow us to observe cycles in which (Selfish, Safe) is chosen over (Selfish, Risky) (Figure 11d).



(a) Bundles selected for analysis

(b) Binary attribute representation

(c) Calibration based on $\underline{R}(P)$

(d) Calibration based on $\overline{R}(P)$

Figure 11: Binary attribute representation of Exley (2016)'s choice data

In panels (c) and (d) bundles are represented by $(Recipient, Prize, Probability)$ where $Recipient = 1$ denotes Charity and $Recipient = 0$ denotes Self. Choice sets marked in orange are observed in the data. Choices labeled (i) follow from the calibration of $\underline{R}(P)$ and $\overline{R}(P)$. Choices labeled (ii) are not directly observed in the data, but are imputed from the calibration of L plus *linearity in payoffs*.

Exley's analysis compares certainty equivalents of different lotteries – i.e., indifference points – measured in terms of payments to self and to charity (see her Prediction 3). She assumes that certainty equivalents are the midpoints of the last-rejected and first-accepted options on the choice lists. Our analysis uses the observed choices only, so is expressed in terms of strict preferences.

B.3.3 Imputing non-observed choices

Figures 11c and 11d include preferences on the upper horizontal choice set. This choice set is not observed in the data. Instead, we use Exley's *linearity in payoffs* assumption once again, to impute what the participant would have chosen if given the choice. Exley's analysis uses this assumption to infer the ranking of certainty equivalents measured in self and charity dollars. We use it to infer the ranking of two specific safe payoffs.

In our constructed binary attribute space all prizes to charity are an $L = {}^{X}\!/_{10}$ multiple of prizes to self. Since we know that $\$X$ to charity was chosen over \$10 to self in the normalization choice list, *linearity in payoffs* implies that $(1, \$10LR(P), 1)$ (a sure payment to charity) would be chosen over $(0, \$10R(P), 1)$ (a sure payment to self) for any value of $R(P)$. Hence we can impute the upper horizontal choice labeled (ii) in the diagrams: (Charity, Safe) $\succ$ (Self, Safe).

Our calibration of the binary attribute space is constrained by the lotteries that we observe, since these always have prizes \$10 and $\$X$, forcing us to fix $L = {}^{X}\!/_{10}$. As a result, we cannot construct a binary attribute space with a different value of L and hence we cannot observe or impute the opposite preference (Charity, Safe) $\prec$ (Self, Safe).

B.3.4 Separable representation

Now that our binary attribute space is defined, we show that the preferences defined by (4) have a separable representation over its attributes.

Consider the four bundles defined by our calibration parameters L and $R(P)$. Using (4) we have:

$$
\begin{aligned}
u\begin{pmatrix} Selfish \\ Risky \end{pmatrix} &= u(0, 10, P) &&= \pi(P)\,(10)^{\alpha} \\
u\begin{pmatrix} Charity \\ Risky \end{pmatrix} &= u(1, 10L, P) &&= \pi(P)\left(\tfrac{10L}{1+\lambda}\right)^{\alpha} \\
u\begin{pmatrix} Selfish \\ Safe \end{pmatrix} &= u(0, 10R(P), 1) &&= (10R(P))^{\alpha} \\
u\begin{pmatrix} Charity \\ Safe \end{pmatrix} &= u(0, 10LR(P), 1) &&= \left(\tfrac{10LR(P)}{1+\lambda}\right)^{\alpha}
\end{aligned}
$$

The ranking of bundles is invariant to any increasing transformation of u. Consider the transformation

$$U(x) := 2\ln\left(\frac{u(x)^{\frac{1}{\alpha}}}{10}\right) - \ln\left(R(P)\pi(P)^{\frac{1}{\alpha}}\right) + \ln\left(\frac{1+\lambda}{L}\right)$$

Applying this, we obtain:

$$\begin{aligned}
U\begin{pmatrix} Selfish \\ Risky \end{pmatrix} &= \ln\left(\tfrac{1+\lambda}{L}\right) + \ln\left(\tfrac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) \\
U\begin{pmatrix} Charity \\ Risky \end{pmatrix} &= -\ln\left(\tfrac{1+\lambda}{L}\right) + \ln\left(\tfrac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) \\
U\begin{pmatrix} Selfish \\ Safe \end{pmatrix} &= \ln\left(\tfrac{1+\lambda}{L}\right) - \ln\left(\tfrac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right) \\
U\begin{pmatrix} Charity \\ Safe \end{pmatrix} &= -\ln\left(\tfrac{1+\lambda}{L}\right) - \ln\left(\tfrac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right)
\end{aligned}$$

Let $Social \in \{Generous, Selfish\}$ be $x_1 \in \{-1, 1\}$ and $Risk \in \{Safe, Risky\}$ be $x_2 \in \{-1, 1\}$. Then we have:

$$U(x) = \ln\left(\frac{1+\lambda}{L}\right)x_1 + \ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right)x_2 \tag{6}$$

Thus, any choices over the four bundles can represented by a separable utility function in x_1 and x_2.

The Signaling foundation requires the observer to have mean-zero Gaussian priors over the weights. Thus, it requires ${}^{1+\lambda}/_{L} \sim \text{Lognormal}(0, \sigma_1^2)$ and ${}^{\pi(P)^{\frac{1}{\alpha}}}/_{R(P)} \sim \text{Lognormal}(0, \sigma_2^2)$. The zero mean requires L and $R(P)$ to be appropriately calibrated to make the participant indifferent in expectation.

B.3.5 Permutation tests

We perform two simple permutation tests that ask whether our data are consistent with different assumptions about noise in behavior.

The starting point is the experimental dataset. An observation is C_{iP} where $i \in (1, ..., 86)$ indexes participants and $P \in (.05, .1, .25, .5, .75, .9, .95)$ indexes lottery probabilities. $C \in \{0, 1, 2, 3\}$ records what type of cycle was observed for that participant-probability: 0 for no cycle, 1 for pro-Risky, 2 for pro-Safe, 3 for pro-Selfish.

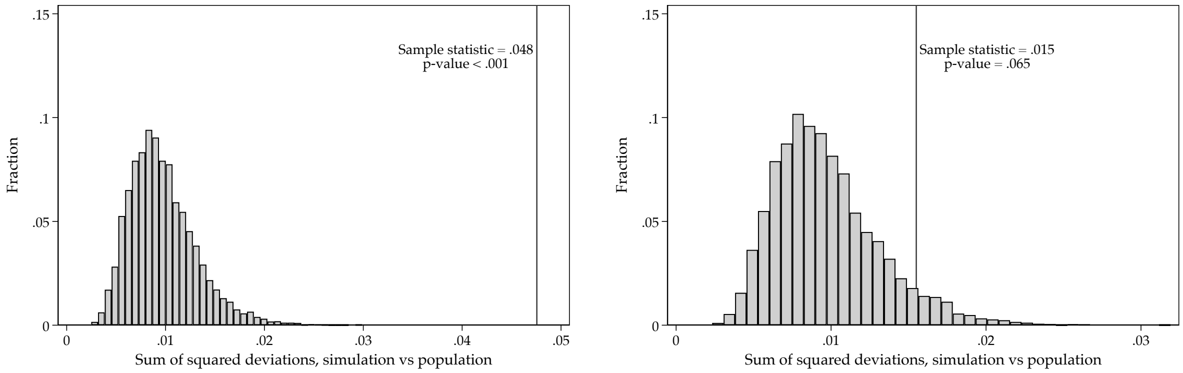


Figure 12: Permutation tests on Exley (2016) data

Left panel corresponds to Null hypothesis 1, right panel to Null hypothesis 2.

Null hypothesis 1: homogeneity. The null hypothesis of our first permutation test is that, separately for each value of P, the likelihood of a given cycle is the same for all participants. E.g., it could be that when $P = 0.5$, all participants have a 5% chance of a pro-Risky cycles, a 10% chance of a pro-Safe, cycle, and an 18% chance of a pro-Selfish cycle. Our permutation test therefore permutes indices i (we do this within each probability P, since cycle probabilities vary across probabilities). Intuitively, this test asks whether variation in cycling behavior could be explained by a representative agent with a certain implicit preference type, plus noise.

Null hypothesis 2: homogeneity conditional on a cycle. The null hypothesis of our second permutation test is that, separately for each value of P, *conditional on a cycle being observed* the likelihood of a given cycle is the same for all participants. E.g., it could be that when $P = 0.5$, 15% of *cycles* are pro-Risky, 30% are pro-Safe, and 55% are pro-Selfish (corresponding to the relative frequencies in the data). Conditional on cycling, these probabilities are the same for all participants. Our permutation test therefore permutes indices i within each probability P *conditional on* $C_{iP} \neq 0$. Intuitively, this test asks whether variation in cycling behavior could be explained by heterogeneity in the *likelihood* of cycling, but otherwise homogeneity in implicit preferences, plus noise.

The basic testing approach is as follows.

1. We represent each participant in the **sample** according to their number of cycles of each type (pro-Risky, pro-Safe, pro-Selfish). We then compute the fraction of participants exhibiting each possible combination. We call these the **sample proportions**. For example, 20 percent of participants have no cycles $(0, 0, 0)$.

2. We duplicate the experimental dataset 10000 times, creating a **population** of 860,000

decision-makers that holds constant the frequency of each observed choice. We then randomly permute rows of this dataset according to our null hypothesis to generated a simulated population distribution of behavior under the null. We compute the fraction of the population exhibiting each possible combination of cycles, and call these the **population proportions**.

3. We compute the sum of squared differences between sample and population proportions, this is our *sample statistic* of interest. A small value of this statistic implies the sample distribution is similar to the population distribution.
4. Returning to the sample dataset with 86 participants, we generate 10,000 **simulated samples**, by permuting rows according to the null assumptions. For each, we compute the fraction of the simulated sample exhibiting each combination of cycles, and call these the **simulated proportions**. We compute the sum of squared differences between the simulated proportions and the population proportions, to obtain a 10,000 draws of the **simulated statistic**.
5. The p-value of the test is simply the fraction of simulated statistics that are larger than the sample statistic. A small p-value indicates that the sample statistic tends to be larger than we would expect it to be under the null hypothesis.

We present our findings in Figure 12. We strongly reject Null hypothesis 1 ($p < .001$). The main contributor to this rejection is substantial excess mass at $(0, 0, 0)$ in the sample relative to that expected under the null: 20 percent of participants have no cycles at all, whereas under the null only around 5 percent of participants should exhibit zero cycles across all seven probabilities. We also find evidence against Null hypothesis 2 ($p = .065$), albeit weaker.