Implicit Preferences[*]

Tom Cunningham[†]   Jonathan de Quidt[‡]

November 15, 2022

Latest version available 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 the more the attribute is mixed with other attributes ("dilution"). We show that implicit preferences arise under a diverse set of psychological 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 from joint evaluation data. We apply the model to two published datasets, finding evidence for implicit risk preferences, implicit selfishness, and implicit discrimination.

JEL codes: D91, J71, C90, D83

---

1 Introduction

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

Inconsistencies in decision-making are often described as arising from a conflict between opposing motives. In this paper we a formalize a common intuition about how motives interact and fully characterize its testable implications. Our theory is consistent with a variety of psychological foundations for the underlying conflict, and easy to apply empirically.

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:
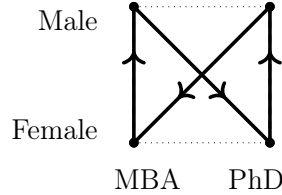
The choices are intransitive and therefore inconsistent with standard utility maximization. Nevertheless they form an intuitive "figure 8" pattern, suggesting two distinct attitudes towards gender: favoring women when the candidates differ only in gender, but favoring men when the candidates additionally differ in other respects.

We generalize the observation that decisions can reveal 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 a positive, negative, or neutral implicit preference for each attribute. We identify the direction of implicit preferences from behavior with a "dilution" assumption: the influence of an implicit preference for an attribute increases whenever that attribute is mixed 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 data on *evaluations* such as willingness to pay, teachers' grading decisions, or judges' sentencing decisions, when each evaluation involves a definite 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 a male candidate would be assigned a lower wage when he is compared to a woman with the same qualification, than when he is compared to a woman with a different qualification.

Section 2 presents our formal model. We assume a comparative utility function, whereby the utility of a bundle $\boldsymbol{x}$ (which we will call the "target"), depends on another bundle $\boldsymbol{z}$ (its "comparator"): we write $u(\boldsymbol{x}, \boldsymbol{z})$. We assume that implicit preferences over attributes are separable and that an implicit preference's strength of influence depends on both $\boldsymbol{x}$ and $\boldsymbol{z}$, according to a partial order over comparisons called *influence-dominance*.

Our core result is a representation theorem stating that a set of choices or evaluations has an implicit preferences representation if and only if there does not exist a matching between decisions, such that implicitly-preferred bundles are ranked *higher* in influence-*dominated* decisions, and vice versa. Concretely, if someone implicitly prefers men, we should not observe them choose women over men when the influence of implicit gender preferences is strong, and men over women when the influence of implicit gender preferences is weak.

The theorem is general with respect to the influence-dominance relation. To apply it we add two specific assumptions: (1) "Equivalence": the influence of implicit preferences depends only on which attributes are shared or not shared between target and comparator ($|\boldsymbol{x}-\boldsymbol{z}|$); and (2) "Dilution": an implicit preference more strongly influences decisions when its attribute is mixed with a superset of other attributes.[1] We also describe a third assumption useful for some analyses of evaluation data, effectively assuming implicit preferences have greater influence when an attribute is shared between $\boldsymbol{x}$ and $\boldsymbol{z}$ than when it is not. All three assumptions are consistent with the psychological foundations that we introduce later.

Given these assumptions, the "figure 8" pattern illustrated above identifies an implicit preference for men over women. Section 3 describes a number of other intuitive patterns in choice ("right triangle," "parallel triangles," "square") and in evaluation ("scissor," "parallel scissor"), and shows what they reveal about the decision maker's implicit preferences. The examples we provide are easy to test for in empirical applications.

Our definition of "implicitness" is thus behavioral, similar to decision-theoretic concepts like complementarity or elasticity which are defined without reference to the underlying psychology. However, we believe it captures a core intuition found in multiple distinct literatures. To demonstrate this, Section 4 describes three distinct maximizing models,

---

[1]Example: Suppose the target is a female with a Harvard PhD. Let Comparison 1 compare her to a male Harvard PhD, and Comparison 2 compare her to a male Harvard *MBA*. Dilution says that implicit gender preferences have more influence in Comparison 2, because gender is mixed with qualification. Now let Comparison 3 compare her to a male *Yale MBA*. The influence of implicit gender preferences strengthens further, as gender is now mixed with qualification and college. Finally, let Comparison 4 be with a male *Yale PhD*. Dilution does not rank Comparisons 2 and 4, because neither mixture is a superset of the other.

which we call *foundations*, that exhibit implicit preferences in our sense.

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, a special case of models in which case 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)). According to this model the choices above reveal that the hiring manager prefers men over women, but is constrained by a rule penalizing the choice of a man over an equally-qualified women.

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, including self-signaling, excuse-driven behavior and "moral wiggle room" (e.g. Bodner and Prelec (2003); Benabou and Tirole (2003, 2006); Norton et al. (2004); Dana et al. (2006, 2007); Andreoni and Bernheim (2009); Exley (2016); Bursztyn et al. (2022)). According to this model the figure 8 pattern above can be explained by a sincere preference for men joined with a signaling motive to favor women. When choosing between a man and a woman who additionally differ in qualification, the choice is less revealing about gender preference, and so the signaling motives are weakened.

The third foundation is an *implicit associations* decision maker for whom some knowledge is tacit. The model is based on Cunningham (2016) and relates to psychological theories of implicit bias and unconscious judgment (e.g. Devine (1989); Greenwald and Krieger (2006); Greenwald et al. (1998); Kahneman (2011); Rand et al. (2012)). In this model the hiring manager is a composite of two rational agents, each with private information. The pre-conscious brain associates men with high value—it has a "good feeling" about the male candidate—but the conscious brain believes gender-based associations are irrelevant. When candidates differ only on gender, the conscious brain can identify and ignore the pro-male association, choosing the woman. But the more gender is mixed with other attributes, the harder it is to disentangle relevant from irrelevant associations, leading it to choose the man.

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.

Next we turn to applications. Section 5 provides guidance on using our framework in practice, then section 6 applies it to two existing datasets. We find evidence of implicit selfishness and implicit risk attitudes in choice data from Exley (2016), and 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, and Section 7 discusses how these types of models will not generally exhibit implicit preferences in our sense.[2] 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 that the patterns of behavior that we highlight have intuitive appeal.

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 choose from a choice set the element which maximizes their true preference from within the subset which are "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.

An advantage of using bundles of attributes, as we do, is that implicit preferences can be identified from from binary choices.[3] Additionally, linking implicit preferences to attributes instead of atomic outcomes facilitates out-of-sample predictions: our hiring manager's gender bias can be predicted to carry over to choice between new candidates with different characteristics. Finally, our model applies to both choice and evaluation, unlocking a wide range of additional applications. This set of features makes our framework particularly well suited to applied work, and we provide an extensive collection of identification tools that can be implemented in existing datasets and new experiments. We demonstrate this with our own applications, and our tools have recently been adopted by others: Barron et al. (2022) apply our approach and find evidence of implicit gender bias.

There is an extensive psychological literature on implicit attitudes but as far as we know ours is the only definition based on ordinary decision-making. In psychology the term "implicit" is usually applied to cognition, attitudes, judgments, preferences, or knowledge that

---

[2]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 (2016) which identifies a figure 8 pattern in choice is Cubitt et al. (2018), studying intertemporal tradeoffs. See Section 7 for more discussion.

[3]With atomic elements binary choice will generally be uninformative: a cycle of the form $a \succ b \succ c \succ a$ implies that there must exist some constraint on choice, but nothing more.

are "outside conscious attentional focus" (Devine, 1989; 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. Rooth (2010); Reuben et al. (2014); Glover et al. (2017); Alesina et al. (2018); Carlana (2019)). However their interpretation and predictiveness remain controversial (Oswald et al., 2013; Greenwald et al., 2015).

A number of prior empirical studies share the intuition that underlying motives can be revealed by observing comparisons that vary in how direct or transparent they are: Snyder et al. (1979) on discrimination against the disabled, Norton et al. (2004); Uhlmann and Cohen (2005); Bohnet et al. (2016) on gender discrimination, Hodson et al. (2002) on racial discrimination, Caruso et al. (2009) on body weight discrimination, Exley (2016) on excuses for selfish behavior, Cubitt et al. (2018) on time discounting. Each paper uses identification approaches tailored to their setting. We provide a formal foundation for the common intuition, and empirical tools that can be broadly applied.

Our introductory example shows how we can identify implicit discrimination, a topic of great recent interest.[4] There are many other possible applications—this method could in principle 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 might reasonably be anticipated.

2   Model

This section proceeds as follows: (1) We define observable behavior as a "dataset" of inequalities between comparative utilities, written $u(\boldsymbol{x}, \boldsymbol{z}) \gtreqless u(\boldsymbol{x}', \boldsymbol{z}')$. (2) We assume decision makers are endowed with directional, additively separable implicit preferences on each attribute ($\kappa_i \in \{-1, 0, 1\}$). (3) We assume the strength with which an implicit preference

---

[4]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.

The economics literature highlights the distinction between taste-based (Becker, 1957) and statistical (Phelps, 1972; Arrow, 1973) discrimination (of which the latter may be inaccurate: Bohren et al. (2021)). Either type of discrimination can be implicit. Bohren et al. (2022) decompose discrimination into *direct* and *systemic* components. For example direct discrimination early in a woman's career contributes to systemic discrimination later on, as she ends up with a weaker resume than an equally-able man. Our notion of implicit discrimination is a form of direct discrimination; systemic effects can amplify its impact.
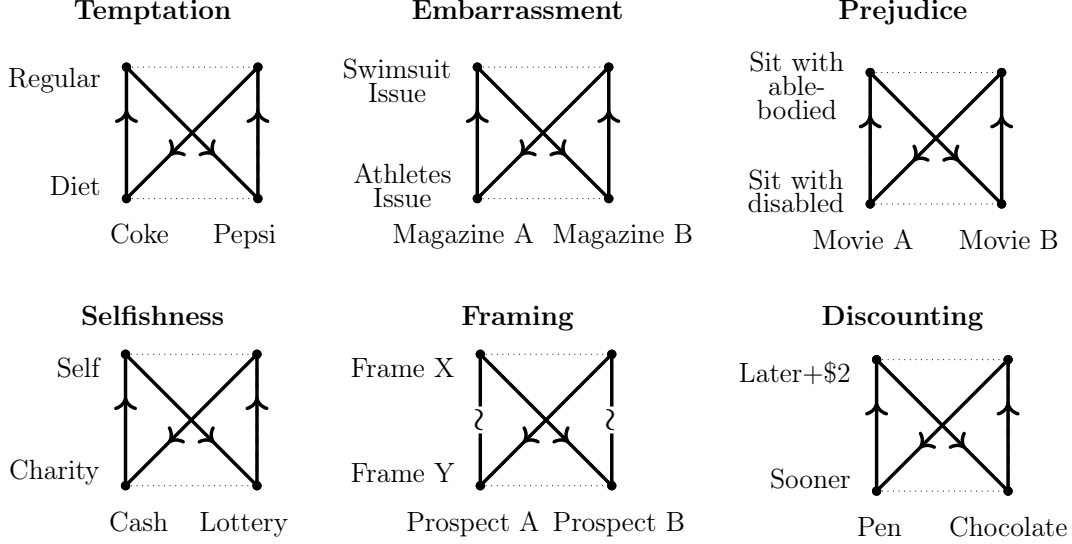
Temptation. The decision maker chooses between diet and full-sugar sodas. They explicitly prefer diet soda, but reveal an implicit preference for the sugary option.

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

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

Selfishness. The decision maker chooses between a lottery and a safe amount, where the beneficiary is themselves or charity. They explicitly prefer to give to charity, but reveal an implicit preference for self. (Inspired by Exley (2016)).

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. They reveal an explicit preference for sooner rewards, but an implicit preference for the delay. (Cubitt et al., 2018).

Figure 1: Figure 8 intransitivities applied to various domains.

influences decisions obeys a given partial order across comparisons, $(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')$. (4) We prove that a dataset admits an implicit preferences *representation* if and only if its inequalities obey a matching condition using the partial orders $\sqsupseteq_i$. (5) We add assumptions on $\sqsupseteq_i$, in particular that implicit preferences have more influence when an attribute is "diluted."

We study preferences over **bundles** of n binary attributes: $\boldsymbol{x} \in \mathcal{X} = \{-1, 1\}^n$, e.g. male vs female, PhD vs MBA, aisle vs window, sugar vs sweetener, sooner vs later.[5,6]

---

[5]Attributes do not need to be intrinsically binary, but our analysis will apply to data containing at most two distinct realizations of each attribute.

[6]All vectors will be column vectors, indicated with a bold font, and $\boldsymbol{x}^T$ will refer to the transpose of $\boldsymbol{x}$. Absolute values of vectors will be element-wise: $|\boldsymbol{x}|^T = \begin{pmatrix} |x_1| & \ldots & |x_n| \end{pmatrix}$. Inequalities between vectors will

We allow the utility of the bundle under consideration, $\boldsymbol{x}$, to depend on a second bundle, $\boldsymbol{z}$, taking the form of a *comparative utility function* $u(\boldsymbol{x}, \boldsymbol{z}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We will refer to $\boldsymbol{x}$ as the *target*, $\boldsymbol{z}$ as the *comparator*, and $(\boldsymbol{x}, \boldsymbol{z})$ as the *comparison*. In principle comparators could come from anywhere but we highlight two types of decision that are inherently comparative: *binary choice* between $\boldsymbol{x}$ and $\boldsymbol{z}$, and *joint evaluation*, where the decision maker simultaneously reports values for $\boldsymbol{x}$ and for $\boldsymbol{z}$ (e.g. willingness to pay).

We treat decisions as revealing inequalities between pairs of utilities.[7] Formally, we define a **dataset** D as a collection of m 4-tuples, $(\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j)_{j=1}^m$, with $\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j \in \mathcal{X}$. Each element of D represents a single inequality: $u(\boldsymbol{x}^j, \boldsymbol{z}^j) \gtreqless u(\boldsymbol{x}'^j, \boldsymbol{z}'^j)$. By convention we order inequalities such that the first $\bar{m}$ are strict (with $\bar{m} \geq 1$), and the remainder are weak:

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

Section 2.2 explains precisely how to construct a dataset from observed choices and evaluations. In short: if we observe that $\boldsymbol{x}$ is chosen over $\boldsymbol{z}$ we treat this as $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{z}, \boldsymbol{x})$. If we observe that $\boldsymbol{x}$ is given a higher evaluation when evaluated alongside $\boldsymbol{z}$ than alongside $\boldsymbol{z}'$ we treat this as $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}, \boldsymbol{z}')$.

We next define an Implicit Preferences utility function, $u^I(\boldsymbol{x}, \boldsymbol{z})$. Utility is the sum of the *explicit value*, a standard utility function $v(\boldsymbol{x})$ that depends only on $\boldsymbol{x}$, and the *implicit value*. The implicit value is a sum over $\boldsymbol{x}$'s attributes, x_i, weighted by (1) the decision maker's implicit preference for that attribute, κ_i; and (2) the *influence* of the implicit preference, $\theta_i(\boldsymbol{x}, \boldsymbol{z})$, which depends on both $\boldsymbol{x}$ and $\boldsymbol{z}$.

$$u^I(\boldsymbol{x}, \boldsymbol{z}) = \underbrace{v(\boldsymbol{x})}_{\text{explicit value}} + \overbrace{\sum_{i=1}^n x_i \underbrace{\kappa_i}_{\substack{\text{implicit} \\ \text{pref for } i}} \underbrace{\theta_i(\boldsymbol{x}, \boldsymbol{z})}_{\substack{\text{influence} \\ \text{of } i}}}^{\text{implicit value}}, \tag{1}$$

with $v : \mathcal{X} \to \mathbb{R}$, $\kappa_i \in \{-1, 0, 1\}$, $\theta_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.[8]

be defined as:

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

[7] This approach allows us to exploit results that give necessary and sufficient conditions for existence of solutions to systems of linear inequalities. Chambers and Echenique (2016) discuss this general approach to decision theory.

[8] Our definition of a dataset only uses ordinal information—inequalities between utilities—so we could without loss of generality (1) wrap $u^I(\boldsymbol{x}, \boldsymbol{z})$ in a strictly increasing function (i.e. linearity is not important), and (2) normalize θ_i to be non-negative.

A decision maker's implicit preferences are either negative, neutral, or positive: $\kappa_i \in \{-1, 0, 1\}$. As the comparator $\boldsymbol{z}$ changes, each implicit preference's influence on utility can increase or decrease. For example, suppose attribute i represents gender, with $x_i = 1$ denoting men, and suppose the decision maker implicitly prefers men ($\kappa_i = 1$). If the influence of implicit preferences on attribute i increases (θ_i increases) then the utility assigned to men will increase and the utility of women will decrease, all else equal.

We encode assumptions on the influence function with a set of partial orders. For each attribute i there exists an *influence-dominance* relation, which is a partial order $\sqsupseteq_i$ over the set of all comparisons $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{X}$. Given two comparisons $(\boldsymbol{x}, \boldsymbol{z})$ and $(\boldsymbol{x}', \boldsymbol{z}')$ we describe $(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')$ as $(\boldsymbol{x}, \boldsymbol{z})$ **influence-dominates** $(\boldsymbol{x}', \boldsymbol{z}')$ **on attribute** i. We then assume that the influence function obeys this partial order:

Definition 1 (Influence-dominance). $(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}') \implies \theta_i(\boldsymbol{x}, \boldsymbol{z}) \geq \theta_i(\boldsymbol{x}', \boldsymbol{z}')$.

If $(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')$ and $(\boldsymbol{x}', \boldsymbol{z}') \sqsupseteq_i (\boldsymbol{x}, \boldsymbol{z})$, it follows that $\theta_i(\boldsymbol{x}, \boldsymbol{z}) = \theta_i(\boldsymbol{x}', \boldsymbol{z}')$.

Our representation theorem takes the set of influence-dominance relations $\sqsupseteq_i, i \in \{1, ..., n\}$ as given, i.e. it holds for any assumptions on influence-dominance. Section 2.1 introduces specific assumptions on $\sqsupseteq_i$, motivated by theory, most importantly the Dilution assumption.

We can now define an Implicit Preferences Representation.

Definition 2 (Implicit Preferences Representation). *A dataset D has an **Implicit Preferences Representation** if and only if there exists an explicit value function $v : \mathcal{X} \to \mathbb{R}$, a vector of implicit preferences $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$, and a set of influence functions $\boldsymbol{\theta} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^n$ such that (1) $u^I(\boldsymbol{x}, \boldsymbol{z})$ satisfies every inequality in D, and (2) $\boldsymbol{\theta}$ obeys Influence-dominance.*

We will also say that a given vector of implicit preferences, $\boldsymbol{\kappa}$, **rationalizes** dataset D if and only if D has an Implicit Preferences Representation with implicit preferences equal to $\boldsymbol{\kappa}$. Our main result, Theorem 1, allows us to characterize all $\boldsymbol{\kappa}$s that can rationalize a dataset.

To state the theorem we need a few additional terms. First, our analysis makes use of weighted subsets of the dataset, which we call *cyclical selections*:

Definition 3 (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 a vector of non-negative integer weights $\boldsymbol{s} \in \mathbb{N}^m$ that select inequalities such that each bundle appears equally often on the left- 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).

A cyclical selection consists of one or more sequences of inequalities that begin and end with the same target ($u(\boldsymbol{x}, .) > ... > u(\boldsymbol{x}, .)$), so cannot be rationalized by standard preferences $u(\boldsymbol{x}, \boldsymbol{z}) = v(\boldsymbol{x})$.[9]

Our theorem will be stated in terms of a set of 1:1 matchings between inequalities in a cyclical selection. Specifically we will be matching "wins" to "losses", defined as follows. Given an inequality $u(\boldsymbol{x}, \boldsymbol{z}) \gtreqless u(\boldsymbol{x}', \boldsymbol{z}')$ we count a **win** for attribute i if the higher-ranked bundle, $\boldsymbol{x}$, has a positive realization of that attribute ($x_i = 1$), and/or the lower-ranked bundle, $\boldsymbol{x}'$, has a negative realization ($x'_i = -1$). We define a **loss** as the reverse. For attribute i and comparison $(\boldsymbol{x}, \boldsymbol{z})$, the **score** $c_{i,(\boldsymbol{x}, \boldsymbol{z})}$ equals total wins minus total losses:[10]

Definition 4 (Score). *Given a dataset $D = \{x^j, z^j, x'^j, z'^j\}_{j=1}^m$ and a cyclical selection $\boldsymbol{s} \in \mathbb{N}^m$ the **score** vector, $\boldsymbol{c} \in \mathbb{Z}^{n|\mathcal{X}|^2}$ represents for each $i \in \{1, \dots, n\}$ and $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{X}$, the net number of times that the attribute wins in $\boldsymbol{s}$:*

$$c_{i,(\boldsymbol{x}, \boldsymbol{z})} = \underbrace{\sum_{j:(\boldsymbol{x}^j, \boldsymbol{z}^j) = (\boldsymbol{x}, \boldsymbol{z})} s_j x_i^j}_{\text{inequalities with } (\boldsymbol{x}, \boldsymbol{z}) \text{ on LHS}} - \underbrace{\sum_{j:(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = (\boldsymbol{x}, \boldsymbol{z})} s_j x_i'^j}_{\text{inequalities with } (\boldsymbol{x}, \boldsymbol{z}) \text{ on RHS}}.$$

In a cyclical selection each bundle appears equally often on the left- and right-hand sides, so for each attribute i the sum of scores must equal zero: $\sum_{\boldsymbol{x}, \boldsymbol{z}} c_{i,(\boldsymbol{x}, \boldsymbol{z})} = 0$.

If $\kappa_i = 1$ then we expect bundles with $x_i = 1$ to be chosen over bundles with $x_i = -1$ relatively more often when the influence of implicit preferences over attribute i is stronger. Concretely, if someone implicitly prefers men, we should not observe a cycle of choices in which women are chosen over men when the influence of implicit gender preferences is strong, and men are chosen over women when the influence of implicit gender preferences is weak. We do not directly observe influence, but we can rank it with $\sqsupseteq_i$. Establishing a contradiction entails finding a 1-1 matching between wins and losses that can be ranked according to influence-dominance. Since wins and losses with the same comparison $(\boldsymbol{x}, \boldsymbol{z})$ have the same influence (so can match to one another), it is enough to check for a matching between net wins and net losses, that is, between scores. We now formally define such matchings:

---

[9]E.g., an $\boldsymbol{s}$ that selects the single inequality $u(\boldsymbol{x}, \boldsymbol{x}') > u(\boldsymbol{x}, \boldsymbol{x}'')$, constitutes a cyclical selection, as does an $\boldsymbol{s}$ that selects the three inequalities $u(\boldsymbol{x}, \boldsymbol{x}') > u(\boldsymbol{x}', \boldsymbol{x})$, $u(\boldsymbol{x}', \boldsymbol{x}'') > u(\boldsymbol{x}'', \boldsymbol{x}')$, $u(\boldsymbol{x}'', \boldsymbol{x}) > u(\boldsymbol{x}, \boldsymbol{x}'')$.

[10]E.g., let Male = 1. We count one win for each man on the left-hand side or woman on the right-hand side, and one loss for each woman on the left-hand side or a man on the right-hand side. An inequality has two wins, two losses, or a win and a loss, per attribute. Total wins minus losses for $i, (\boldsymbol{x}, \boldsymbol{z})$ equal:

$$\sum_{j:(\boldsymbol{x}^j, \boldsymbol{z}^j) = (\boldsymbol{x}, \boldsymbol{z})} s_j \mathbb{1}\{x_i^j = 1\} + \sum_{j:(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = (\boldsymbol{x}, \boldsymbol{z})} s_j \mathbb{1}\{x_i'^j = -1\} - \sum_{j:(\boldsymbol{x}^j, \boldsymbol{z}^j) = (\boldsymbol{x}, \boldsymbol{z})} s_j \mathbb{1}\{x_i^j = -1\} - \sum_{j:(\boldsymbol{x}'^j, \boldsymbol{z}'^j) = (\boldsymbol{x}, \boldsymbol{z})} s_j \mathbb{1}\{x_i'^j = 1\}.$$

The simplified statement of $c_{i,(\boldsymbol{x}, \boldsymbol{z})}$ in Definition 4 exploits $\mathbb{1}\{x_i^j = 1\} - \mathbb{1}\{x_i^j = -1\} = x_i^j$.

Definition 5 (wins influence-dominate losses). *Given a vector of scores for attribute i, $\boldsymbol{c}_i \in \mathbb{Z}^{|\mathcal{X}|^2}$ we say **wins influence-dominate losses for attribute** i if there exists a matrix of non-negative integers $M_i \in \mathbb{N}^{|\mathcal{X}|^2 \times |\mathcal{X}|^2}$ with:*

$$\forall \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}' \in \mathcal{X}, \quad (M_{i,(\boldsymbol{x},\boldsymbol{z}),(\boldsymbol{x}',\boldsymbol{z}')} > 0) \implies (\boldsymbol{x},\boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}',\boldsymbol{z}') \qquad \textit{(matches obey dominance)}$$

$$\forall \bar{\boldsymbol{x}}, \bar{\boldsymbol{z}} \in \mathcal{X}, \quad c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})} = \underbrace{\sum_{\boldsymbol{x},\boldsymbol{z} \in \mathcal{X}} M_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}}),(\boldsymbol{x},\boldsymbol{z})}}_{\textit{outflow: } (\bar{\boldsymbol{x}},\bar{\boldsymbol{z}}) \textit{ dominates}} - \underbrace{\sum_{\boldsymbol{x}',\boldsymbol{z}' \in \mathcal{X}} M_{i,(\boldsymbol{x}',\boldsymbol{z}'),(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})}}_{\textit{inflow: } (\bar{\boldsymbol{x}},\bar{\boldsymbol{z}}) \textit{ dominated}} \qquad \textit{(all scores matched)}$$

The first condition says that $(\boldsymbol{x}, \boldsymbol{z})$ is only matched to $(\boldsymbol{x}', \boldsymbol{z}')$ if $(\boldsymbol{x}, \boldsymbol{z})$ influence-dominates $(\boldsymbol{x}', \boldsymbol{z}')$. The second condition checks that all scores are matched: each positive score, $c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})} > 0$ (where wins exceed losses) has a net outflow equal to $c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})}$, and each negative score, $c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})} < 0$ has a net inflow equal to $|c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})}|$. Thus M_i can be thought of as a matching between scores in $\boldsymbol{c}_i$.[11] It may be that no matching exists, for example when some comparison in the dataset is not related to any other by $\sqsupseteq_i$.

We likewise say that **losses influence-dominate wins for attribute** i if there is an M_i that satisfies the same conditions, but the last line in the definition sums to $-c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})}$ instead of $c_{i,(\bar{\boldsymbol{x}},\bar{\boldsymbol{z}})}$.

We can now show that the consistency of a dataset D with a vector of implicit preferences, $\boldsymbol{\kappa}$, depends on the existence of a matching between wins and losses for each attribute.

Theorem 1 (Rationalization by $\boldsymbol{\kappa}$). *A vector of implicit preferences $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$ rationalizes dataset D if and only if there exists no cyclical selection $\boldsymbol{s}$ such that, (1) for every attribute with a positive implicit preference ($\kappa_i = 1$), losses influence-dominate wins, and (2) for every attribute with a negative implicit preference ($\kappa_i = -1$), wins influence-dominate losses.*

The proof is given in Section 9. We first show that the dataset and the influence-dominance relations can be expressed as a system of inequalities in matrix form. Rationalizability requires there exist vectors of explicit values $\boldsymbol{v}$, and influences $\boldsymbol{\theta}$, that solve the system. Motzkin's Rational Transposition Theorem (Border, 2013), tells us that a solution exists if and only if there is no weighting of the rows in the matrix that sums to zero. Finally, we show that existence of the weighted sum is equivalent to our matching condition.

The set of implicit preferences that can rationalize the dataset are those not ruled out by Theorem 1. To verify that a given $\boldsymbol{\kappa}$ can rationalize D one must in principle check *all*

---

[11]**Worked example.** Let $\boldsymbol{x} = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$, $\boldsymbol{z} = \left(\begin{smallmatrix}-1\\-1\end{smallmatrix}\right)$, and $\boldsymbol{z}' = \left(\begin{smallmatrix}-1\\1\end{smallmatrix}\right)$. Consider a cyclical selection with one inequality: $u(\boldsymbol{x},\boldsymbol{z}) > u(\boldsymbol{x},\boldsymbol{z}')$. Let us assume $(\boldsymbol{x},\boldsymbol{z}) \sqsupseteq_1 (\boldsymbol{x},\boldsymbol{z}')$. We have $c_{1,(\boldsymbol{x},\boldsymbol{z})} = 1$ (a single win) and $c_{1,(\boldsymbol{x},\boldsymbol{z}')} = -1$ (a single loss). There exists a matching matrix for attribute 1 that matches the positive score on $(\boldsymbol{x},\boldsymbol{z})$ to the negative score on $(\boldsymbol{x},\boldsymbol{z}')$: $M_1 = \left[\begin{smallmatrix}0 & 1\\0 & 0\end{smallmatrix}\right]$. Hence, wins influence-dominate losses for attribute 1.

cyclical selections in D. For simple datasets this is usually straightforward and may be possible by visual inspection. For larger datasets the search may be simplified by using the matrix representation of the problem. The next Corollaries follow immediately:

Corollary 1 (Representation). *A dataset D has an Implicit Preferences Representation if and only if there exists at least one $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$ satisfying the conditions of Theorem 1.*

Corollary 2 (Rationalization by standard preferences). *A dataset D can be rationalized by a standard utility function (i.e. $\boldsymbol{\kappa} = \boldsymbol{0}$) if and only if it does not contain a cyclical selection.*

Proof: if $\boldsymbol{\kappa} = \boldsymbol{0}$ the matching condition is trivially satisfied for all i in any cyclical selection.

Our theory is falsified if the data cannot be rationalized by any $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$. We have not found a necessary and sufficient condition for non-rationalizability simpler than checking every $\boldsymbol{\kappa}$. But there is a simple sufficient condition for a dataset to be non-rationalizable:

Corollary 3 (Falsification). *A dataset D has no Implicit Preferences Representation if it contains a cyclical selection $\boldsymbol{s}$ such that for every attribute, losses influence-dominate wins and wins influence-dominate losses.*

This arises when every win within the cyclical selection can be matched to a loss with equal influence. No $\boldsymbol{\kappa}$ can rationalize such a pattern.[12] We provide examples in Section 3.

2.1 Assumptions on Influence

We now add assumptions on the influence-dominance relations ($\sqsupseteq_i$) to tailor our representation theorem to applications. Our assumptions are all motivated by the foundational models that we present in section 4. However we note that Theorem 1 is more general, and holds for any set of partial orders $\sqsupseteq_i, i = \{1, ..., n\}$.

In general, influence $\theta_i(\boldsymbol{x}, \boldsymbol{z})$ can take $|\mathcal{X}|^2 = 2^{2n}$ unique values, one per comparison $(\boldsymbol{x}, \boldsymbol{z})$. Our first assumption, Equivalence, shrinks this set. We assume that influence depends only on which attributes are *shared* ($|x_i - z_i| = 0$) and which are *non-shared* ($|x_i - z_i| = 2$).

Assumption 1 (Equivalence). *For any $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}' \in \mathcal{X}$:*

$$|\boldsymbol{x} - \boldsymbol{z}| = |\boldsymbol{x}' - \boldsymbol{z}'| \implies (\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}') \text{ and } (\boldsymbol{x}', \boldsymbol{z}') \sqsupseteq_i (\boldsymbol{x}, \boldsymbol{z}), \forall i$$

---

[12]The condition in Corollary 3 is sufficient because it implies a single cyclical selection that rules out every possible $\boldsymbol{\kappa}$ vector. But it is not necessary: one can construct examples where no single cyclical selection falsifies the model, but multiple cyclical selections exist that collectively do. See Web Appendix A.1.1.

Equivalence means that influence is identical between any two comparisons with the same sets of shared and non-shared attributes. For example, $\theta_i \left(\left(\begin{smallmatrix} \text{Male} \\ \text{MBA} \end{smallmatrix} \right), \left(\begin{smallmatrix} \text{Female} \\ \text{PhD} \end{smallmatrix} \right) \right) = \theta_i \left(\left(\begin{smallmatrix} \text{Female} \\ \text{MBA} \end{smallmatrix} \right), \left(\begin{smallmatrix} \text{Male} \\ \text{PhD} \end{smallmatrix} \right) \right)$. When we refer to the **status** of an attribute we will mean whether it is shared or non-shared.

Equivalence is a powerful assumption because $|\boldsymbol{x} - \boldsymbol{z}| \in \{0, 2\}^n$, implying $\theta_i(\boldsymbol{x}, \boldsymbol{z})$ can take at most 2^n unique values. Our hiring manager example has just two types of comparison: the vertical choice sets with $|\boldsymbol{x} - \boldsymbol{z}| = \left[\begin{smallmatrix} 0 \\ 2 \end{smallmatrix} \right]$, and the diagonals with $|\boldsymbol{x} - \boldsymbol{z}| = \left[\begin{smallmatrix} 2 \\ 2 \end{smallmatrix} \right]$. Thus there are four choices but at most two unique θ_i values per attribute. This makes the figure 8 cycle a powerful tool for identifying implicit preferences.[13]

Equivalence implies $\theta_i(\boldsymbol{x}, \boldsymbol{z}) = \theta_i(\boldsymbol{z}, \boldsymbol{x})$. This sharpens the inferences we can draw from choice data, because it implies implicit preferences on shared attributes are irrelevant for choice. To see this, observe that $u^I(\boldsymbol{x}, \boldsymbol{z}) - u^I(\boldsymbol{z}, \boldsymbol{x}) = v(\boldsymbol{x}) - v(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\kappa_i \theta_i(\boldsymbol{x}, \boldsymbol{z})$. Shared attributes drop out because $x_i - z_i = 0$.

Our key assumption for identifying the *direction* of implicit preferences is that κ_i's influence increases as additional attributes share status with i. We call it Dilution:

Assumption 2 (Dilution). *For all $i \in \{1, \ldots, n\}$, $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}' \in \mathcal{X}$, with $\delta_i := |x_i - z_i|$*

$$\underbrace{(|x_i' - z_i'| = \delta_i)}_{\substack{i \text{ has same status} \\ \text{in } (\boldsymbol{x}, \boldsymbol{z}) \text{ and } (\boldsymbol{x}', \boldsymbol{z}')}} \wedge \underbrace{\{j : |x_j - z_j| = \delta_i\} \supseteq \{j : |x_j' - z_j'| = \delta_i\}}_{\substack{\text{a superset of attributes share status with } i \\ \text{in } (\boldsymbol{x}, \boldsymbol{z}) \text{ relative to } (\boldsymbol{x}', \boldsymbol{z}')}} \implies \underbrace{(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')}_{\substack{(\boldsymbol{x}, \boldsymbol{z}) \text{ influence-dominates} \\ (\boldsymbol{x}', \boldsymbol{z}') \text{ on } i}}$$

Suppose $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on gender. Dilution implies that an implicit preference favoring one gender will have weak influence when $\boldsymbol{x}$ and $\boldsymbol{z}$ differ *only* on gender, becoming progressively stronger as $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on other attributes in addition to gender. Thus, in our hiring manager example, implicit gender preferences have more influence in the diagonal choice sets than in the vertical choice sets.

Assumptions 1 and 2 are sufficient for all of our analysis of choice data and a number of identification results in evaluation data. In some evaluation examples, our conclusions will depend on how the influence of an implicit preference κ_i changes when i changes status (from shared to non-shared or vice versa). Dilution has nothing to say in such cases, but there are often reasons to think that influence systematically varies with status. Our final assumption formalizes this by designating a particular attribute k as "special," in the sense that the influence of i is always greater when i has the same status as k:

---

[13]Our signaling-choice foundation in Section 4 assumes a naïve observer, which implies Equivalence. A sophisticated observer would adjust for the expected direction of signaling, which violates Equivalence.

Assumption 3 (Dominance of attribute k). *For all $i \in \{1, \ldots, n\} \setminus k$, $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}' \in \mathcal{X}$,*

$$\underbrace{(|x_i - z_i| = |x_k - z_k|) \wedge (|x_i' - z_i'| \neq |x_k' - z_k'|)}_{\substack{i \text{ has same status as } k \text{ in } (\boldsymbol{x}, \boldsymbol{z}) \\ \text{different status from } k \text{ in } (\boldsymbol{x}', \boldsymbol{z}')}} \implies \underbrace{(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')}_{(\boldsymbol{x}, \boldsymbol{z}) \text{ influence-dominates } (\boldsymbol{x}', \boldsymbol{z}')}$$

We think in most cases it is natural to think of k as a shared attribute, capturing what is "held constant" across comparisons. 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 that share status with k.
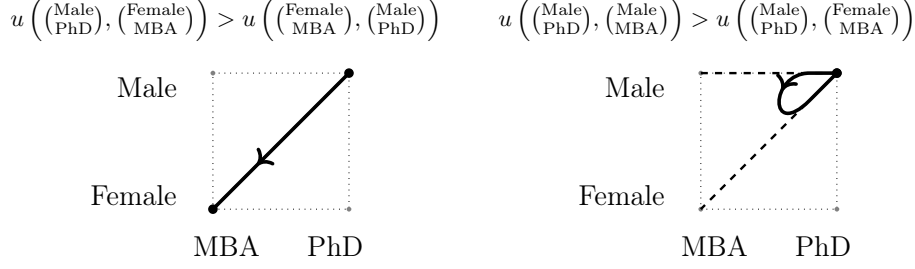
2.2   Choice and Evaluation as Datasets

We have defined a dataset as a set of inequalities between comparative utilities. We now explain precisely how observations from choice and evaluation can be encoded in this format.

For binary choice we treat each bundle as the other bundle's comparator, thus a strict revealed preference for $\boldsymbol{x}$ over $\boldsymbol{z}$ implies a strict inequality, $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{z}, \boldsymbol{x})$, and indifference implies a pair of weak inequalities, $u(\boldsymbol{x}, \boldsymbol{z}) \geq u(\boldsymbol{z}, \boldsymbol{x})$ and $u(\boldsymbol{z}, \boldsymbol{x}) \geq u(\boldsymbol{x}, \boldsymbol{z})$. A choice cycle $\boldsymbol{x} \succ \boldsymbol{x}' \succ \boldsymbol{x}'' \succeq \boldsymbol{x}$ would correspond to a dataset with $m = 4, \bar{m} = 2$: $u(\boldsymbol{x}, \boldsymbol{x}') > u(\boldsymbol{x}', \boldsymbol{x})$; $u(\boldsymbol{x}', \boldsymbol{x}'') > u(\boldsymbol{x}'', \boldsymbol{x}')$; $u(\boldsymbol{x}'', \boldsymbol{x}) \geq u(\boldsymbol{x}, \boldsymbol{x}'')$; and $u(\boldsymbol{x}, \boldsymbol{x}'') \geq u(\boldsymbol{x}'', \boldsymbol{x})$.

We can encode data on continuous *evaluations* of bundles when each evaluated bundle has a clear comparator. For example, when two bundles are evaluated simultaneously we can treat each as the other's comparator. We assume that evaluations are strictly increasing in utility: $y(\boldsymbol{x}, \boldsymbol{z}) = f(u(\boldsymbol{x}, \boldsymbol{z})), f' > 0$. We can then construct a set of inequalities sufficient to represent the ordinal relationships between evaluations. We first rank each evaluation (breaking ties arbitrarily), then enter an inequality for each pair of consecutive evaluations. When two evaluations are equal we use two opposing weak inequalities. For example suppose we observe the following joint evaluations of willingness to pay: $y(\boldsymbol{x}, \boldsymbol{x}') = \$310$, $y(\boldsymbol{x}', \boldsymbol{x}) = \$200$, $y(\boldsymbol{x}, \boldsymbol{x}'') = \$200$, $y(\boldsymbol{x}'', \boldsymbol{x}) = \$150$. Then we would construct a dataset with $m = 4, \bar{m} = 2$: $u(\boldsymbol{x}, \boldsymbol{x}') > u(\boldsymbol{x}', \boldsymbol{x})$, $u(\boldsymbol{x}', \boldsymbol{x}) \geq u(\boldsymbol{x}, \boldsymbol{x}'')$, $u(\boldsymbol{x}, \boldsymbol{x}'') \geq u(\boldsymbol{x}', \boldsymbol{x})$, $u(\boldsymbol{x}, \boldsymbol{x}'') > u(\boldsymbol{x}'', \boldsymbol{x})$.

Graphical representation of inequalities. 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 indicate the location of the two comparators: the dashed lines run from $\boldsymbol{x}$ to $\boldsymbol{z}$ and from $\boldsymbol{x}'$ to $\boldsymbol{z}'$. We give two examples below, where the first visualizes a single choice (the two dashed lines are obscured by the solid line) and the second visualizes an inequality between two evaluations of the same bundle with different comparators.

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

3 Canonical Examples

We now define a set of important classes of dataset, from both choice and evaluation, picked to encompass those that are most useful for applications. We derive the implications of each as a corollary of Theorem 1. The proofs are mechanical so we consign them to web appendix A.1. All definitions are stated in terms of strict inequalities, however it is sufficient for each of these results if at least one inequality in each cyclical selection is strict.

We assume throughout that the influence-dominance relation satisfies Equivalence and Dilution (Assumptions 1 and 2). For evaluation examples we first derive their implications without assuming Dominance of Attribute k (Assumption 3), and then show how adding it refines the implications.

For choice, a cyclical selection consists of one or more intransitive cycles over target bundles, of the form $\boldsymbol{x} \succ \ldots \succ \boldsymbol{x}$. We begin with the *right triangle*, the shortest cycle (three choices) that rules out some implicit preferences. It yields a disjunction over the implicit preferences for all non-shared attributes in the cycle. Second, we discuss the *figure 8*, a four-choice cycle that can unambiguously identify a single implicit preference. Third, we show how pairs of *parallel right triangles* can refine identification relative to the single triangle.

For evaluation, a cyclical selection consists of one or more single inequalities with the same target on the left- and right-hand sides, of the form $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}, \boldsymbol{z}')$.[14] We begin with the *convex scissor*, a single inequality that yields a disjunction over the implicit preferences on *every* attribute. Then we show how pairs of *parallel convex scissors* can refine identification.

Finally we present two examples that imply the existence of some implicit preference but nothing more, and three examples that, if observed, would falsify our assumptions.

For each corollary we give a few concrete examples, in three dimensions. Attribute 1 is

---

[14]Because we construct evaluation datasets by ranking evaluations from highest to lowest (see section 2.2), this inequality may not literally appear in the dataset. Instead we might have $u(\boldsymbol{x}, \boldsymbol{z}) > u(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}})$ in one row and $u(\bar{\boldsymbol{x}}', \bar{\boldsymbol{z}}') > u(\boldsymbol{x}, \boldsymbol{z}')$ in another. But we can construct the intended cyclical selection by including every inequality that lies in between, giving us the sequence $u(\boldsymbol{x}, \boldsymbol{z}) > \ldots > u(\boldsymbol{x}, \boldsymbol{z}')$. This is equivalent to the single inequality, because every intermediate evaluation in the sequence appears on the RHS of one inequality and the LHS of the next, so their wins and losses exactly cancel.
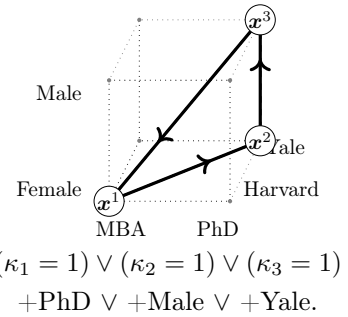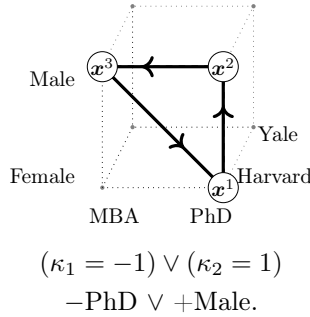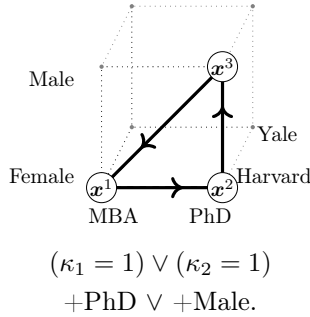
always qualification (PhD $= 1$), attribute 2 is gender (Male $= 1$), attribute 3 is college (Yale $= 1$). We state each example's implications for $\boldsymbol{\kappa}$ and in natural language. E.g., $\kappa_1 = 1$ means we learn κ_1 is positive, $\kappa_1 \neq 0$ means we learn there is an implicit preference for attribute 1 but not its 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 $\pm$Male means we learn there is an implicit gender preference but not its sign.

Choice Examples.

Definition 6 (Right triangle). *A right triangle is a choice cycle over three distinct bundles, ordered $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$, in which $(\boldsymbol{x}^1, \boldsymbol{x}^2)$ and $(\boldsymbol{x}^2, \boldsymbol{x}^3)$ differ on distinct sets of attributes (i.e., $|\boldsymbol{x}^1 - \boldsymbol{x}^2|$ and $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$ are orthogonal).*

Corollary 4. *A right triangle implies at least one nonzero implicit preference favoring $\boldsymbol{x}^3$'s realization of an attribute on which it differs from $\boldsymbol{x}^1$:*

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

$(\kappa_1 = 1) \vee (\kappa_2 = 1)$
$+$PhD $\vee$ $+$Male.

$(\kappa_1 = -1) \vee (\kappa_2 = 1)$
$-$PhD $\vee$ $+$Male.

$(\kappa_1 = 1) \vee (\kappa_2 = 1) \vee (\kappa_3 = 1)$
$+$PhD $\vee$ $+$Male $\vee$ $+$Yale.

A single right triangle cannot unambiguously identify a single implicit preference, because by construction, $\boldsymbol{x}^3$ and $\boldsymbol{x}^1$ must differ on at least two attributes.

In the first example, the matching works as follows: (1) a female MBA is chosen over a female PhD, but rejected in favor of a male PhD (a dilution of the qualification attribute); (2) a male PhD is rejected in favor of a female PhD, but chosen over a female MBA (a dilution of the gender attribute). In both cases, wins influence-dominate losses, so the dataset *cannot* be rationalized by $(\kappa_1 \leq 0) \wedge (\kappa_2 \leq 0)$. Hence we obtain a disjunction: there must be at least one implicit preference, favoring men, favoring PhDs, or both.

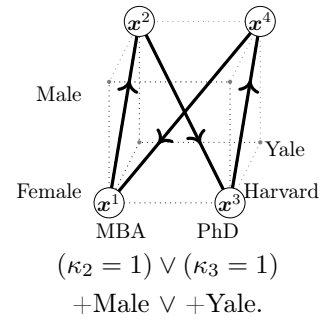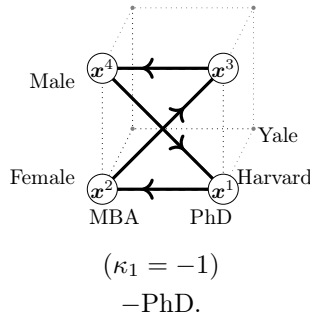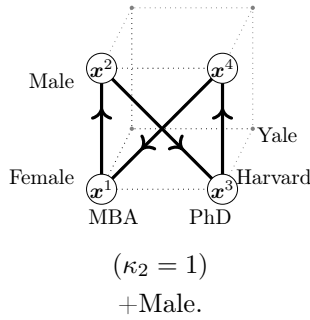**Definition 7** (Figure 8). *A figure 8 is a choice cycle over four distinct bundles, ordered $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^4 \succ \boldsymbol{x}^1$. It must satisfy two conditions: (1) there are only two unique*

vectors of differences $|\boldsymbol{x}^1 - \boldsymbol{x}^2| = |\boldsymbol{x}^3 - \boldsymbol{x}^4|$ *and* $|\boldsymbol{x}^2 - \boldsymbol{x}^3| = |\boldsymbol{x}^1 - \boldsymbol{x}^4|$; *and (2) the latter comparisons differ on a superset of attributes:* $|\boldsymbol{x}^2 - \boldsymbol{x}^3| > |\boldsymbol{x}^1 - \boldsymbol{x}^2|$.

Corollary 5. *A figure 8 implies at least one nonzero implicit preference, favoring* $\boldsymbol{x}^4$*'s realization of an attribute on which it differs from* $\boldsymbol{x}^3$:

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

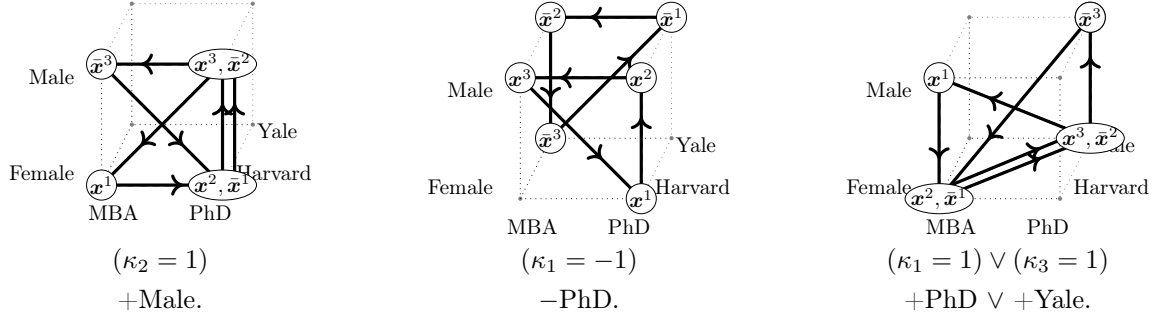| $(\kappa_2 = 1)$ | $(\kappa_1 = -1)$ | $(\kappa_2 = 1) \vee (\kappa_3 = 1)$ |
| +Male. | −PhD. | +Male ∨ +Yale. |

When $\boldsymbol{x}^3$ and $\boldsymbol{x}^4$ differ on a single attribute, we identify the existence of an implicit preference on that attribute and its direction.

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. An implicit preference on the qualification dimension cannot explain both, so there must be an implicit preference favoring men.

Definition 8 (Parallel right triangles)**.** *A pair of parallel right triangles is a cyclical selection consisting of two right triangles* $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$ *and* $\bar{\boldsymbol{x}}^1 \succ \bar{\boldsymbol{x}}^2 \succ \bar{\boldsymbol{x}}^3 \succ \bar{\boldsymbol{x}}^1$, *satisfying two conditions: (1) identical signed differences on* $(\boldsymbol{x}^2, \boldsymbol{x}^3)$ *and* $(\bar{\boldsymbol{x}}^1, \bar{\boldsymbol{x}}^2)$ *(that is,* $\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2$*); and (2) opposing signed differences on* $(\boldsymbol{x}^1, \boldsymbol{x}^2)$ *and* $(\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3)$ *(that is,* $\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3)$*).*

Corollary 6. *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* $\boldsymbol{x}^2$:

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

$(\kappa_2 = 1)$ $\quad$ $(\kappa_1 = -1)$ $\quad$ $(\kappa_1 = 1) \lor (\kappa_3 = 1)$

+Male. $\quad\quad\quad$ −PhD. $\quad\quad\quad$ +PhD $\lor$ +Yale.

If $\boldsymbol{x}^3$ and $\boldsymbol{x}^2$ differ on only one attribute then we can infer that the decision-maker has an implicit preference regarding that attribute and we can infer its sign (in contrast to the disjunction inferred from a single right triangle). Parallel triangles achieve this by eliminating part of each individual triangle's disjunctions. In particular, they eliminate attributes that are non-shared in $(\boldsymbol{x}^1, \boldsymbol{x}^2)$ and $(\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3)$ (where the triangles disagree), leaving only the attributes that are non-shared in $(\boldsymbol{x}^2, \boldsymbol{x}^3)$ and $(\bar{\boldsymbol{x}}^1, \bar{\boldsymbol{x}}^2)$ (where they agree). For instance, the first example above consists of two right triangles that agree on the gender attribute, and disagree on the qualification attribute, allowing us to isolate the implicit gender preference.

Examples without direct comparisons. So far the examples that unambiguously identify a single implicit preference have all included "direct" comparisons with a single non-shared attribute (e.g., between otherwise identical men and women). However a direct comparison is not necessary: it is possible to observe multiple cycles that have no direct comparisons but collectively rule out all but one implicit preference.[15]

Evaluation Examples. We now turn to evaluation data. In each case we first state what can be derived using only Assumptions 1 and 2. Unlike with choice, we cannot ignore implicit preferences on shared attributes, because the left- and right-hand sides of the inequality can have different θ_is. This has two implications. First, in general we identify disjunctions over implicit preferences on *every* attribute. Second, because Dilution does not restrict how influence changes when an attribute goes from shared to non-shared, we sometimes draw indeterminate conclusions about some implicit preferences. We show how Assumption 3 can resolve such indeterminacies.

Definition 9 (Convex scissor). *A convex scissor is a pair of evaluations of a single bundle* $\boldsymbol{x}$ *with two different comparators:* $y^1 = y(\boldsymbol{x}, \boldsymbol{z}^1), y^2 = y(\boldsymbol{x}, \boldsymbol{z}^2)$. *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* $(|\boldsymbol{x} - \boldsymbol{z}^2| > |\boldsymbol{x} - \boldsymbol{z}^1|)$.

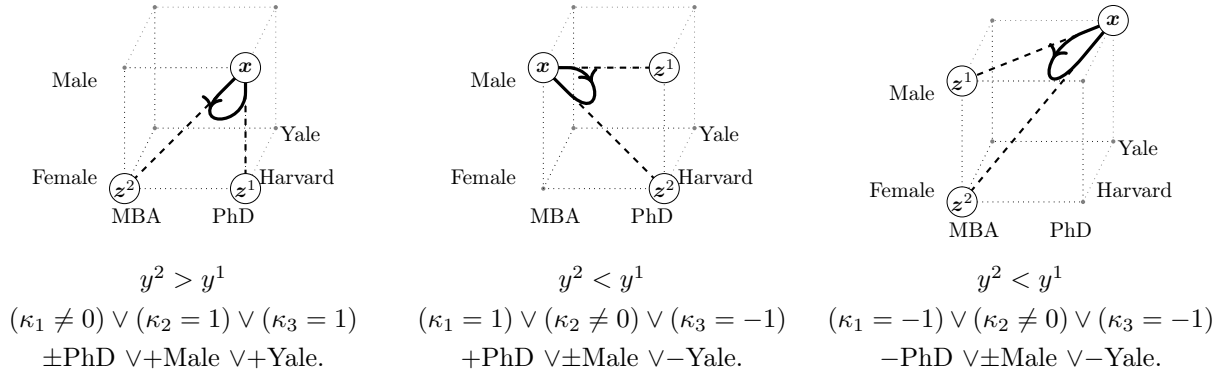**Corollary 7.** *A convex scissor implies at least one nonzero implicit preference:*

---

[15]E.g., our third figure-8 example above reveals $(\kappa_2 = 1) \lor (\kappa_3 = 1)$. If we observed another figure-8 revealing $(\kappa_2 = 1) \lor (\kappa_3 = -1)$, we could conclude that $\kappa_2 = 1$.

$y^2 > y^1$ *(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 indeterminate sign on any other attribute.

$y^2 < y^1$ *(implies the reverse of $y^2 > y^1$)*

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

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

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

The shift of comparison from $\boldsymbol{z}^1$ to $\boldsymbol{z}^2$ changes influence for every attribute. Attributes 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. Attributes that are shared in both comparisons become *less* dilute, because the set of shared attributes shrinks, so their implicit preferences have *less* influence. Attributes that are shared in $|\boldsymbol{x} - \boldsymbol{z}^1|$ but non-shared in $|\boldsymbol{x} - \boldsymbol{z}^2|$ are not restricted by Assumption 2 (Dilution), so we cannot sign their implicit preferences. Assumption 3 (Dominance of attribute k) resolves the ambiguity:

Corollary 8 (Convex scissor with Dominance of attribute k). *Suppose Assumption 3 holds. Let $\Theta = 1$ when k is shared ($x_k = z_k^1 = z_k^2$), and $\Theta = -1$ when k is non-shared ($x_k \neq z_k^1$ and $x_k \neq z_k^2$). A convex scissor implies:*

$$\bigvee_{i:x_i \neq z_i^1} (x_i \kappa_i \Upsilon = 1) \vee \bigvee_{i:x_i = z_i^2} (x_i \kappa_i \Upsilon = -1) \vee \bigvee_{i:z_i^1 \neq z_i^2} (x_i \kappa_i \Upsilon = -\Theta),$$
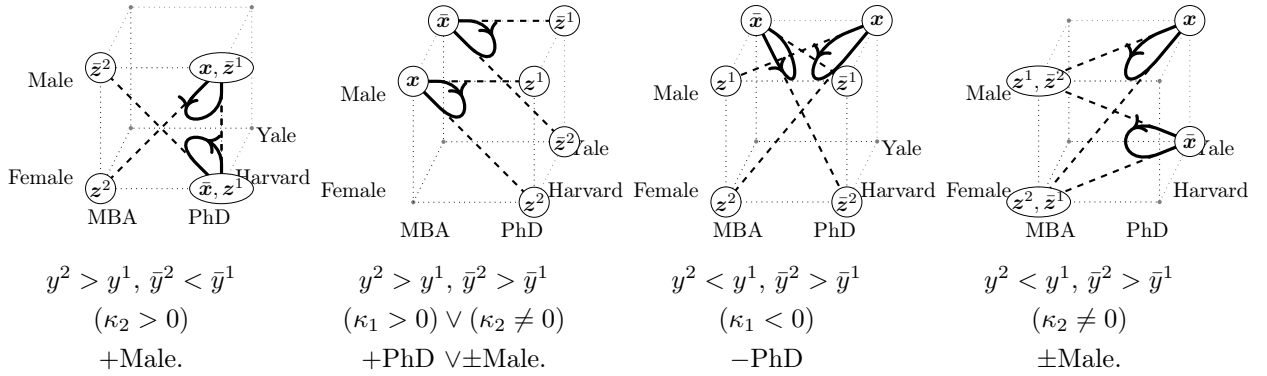
Next, we show that combining two scissors which are reflections of each other can refine our identification of implicit preferences.

Definition 10 (Parallel convex scissors). *A pair of parallel convex scissors is a dataset consisting of two convex scissors, $y^1 = y(\boldsymbol{x}, \boldsymbol{z}^1), y^2 = y(\boldsymbol{x}, \boldsymbol{z}^2)$ and $\bar{y}^1 = y(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^1), \bar{y}^2 =$*

$y(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}^2)$, $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$. *Denote the signs of evaluation changes by* $\Upsilon = sgn(y^2 - y^1)$ *and* $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^1)$. *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.,* $|\boldsymbol{x} - \boldsymbol{z}^2| = |\bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}^2|$*).*[16]

Corollary 9. *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 are summarized in the following disjunction:*

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

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

Just like the pair of parallel right triangles, parallel convex scissors refine the implications of their constituent scissors. This occurs when there are attributes with $x_i\Upsilon = -\bar{x}_i\bar{\Upsilon}$, in which the terms associated with those attributes drop out of the disjunction. Intuitively, the observed behavior cannot be explained by implicit preferences on those attributes if evaluation moves in contradictory directions in the two scissors.

We can unambiguously identify an attribute's implicit preference if we can eliminate all other terms in the disjunction. The first and third examples show that Assumption 3 is not required to accomplish this. In the fourth example, we learn there must be a nonzero implicit preference on attribute 2, but Assumption 3 is needed to learn its sign.

Corollary 10 (Parallel convex scissors with Dominance of attribute k). *Suppose Assumption 3 holds. Let* $\Theta = 1$ *when k is shared* $(x_k = z_k^1 = z_k^2)$ *and* $(\bar{x}_k = \bar{z}_k^1 = \bar{z}_k^2)$, *and* $\Theta = -1$ *when k*

---

[16]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 do not exploit the ranking of evaluations *between* scissors. In principle such information could be used to extract additional information, but we do not model this for sake of brevity.
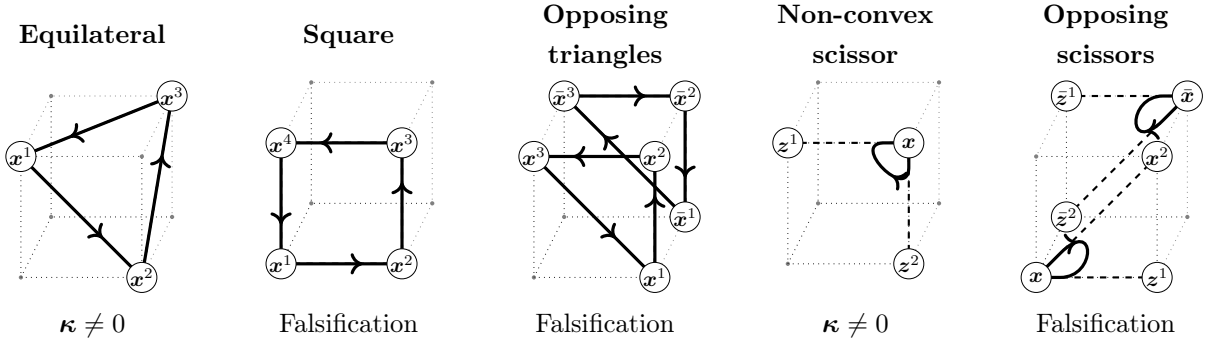
is non-shared ($x_k \neq z_k^1 = z_k^2$) and ($\bar{x}_k \neq \bar{z}_k^1 = \bar{z}_k^2$). A pair of parallel convex scissors implies:

$$\bigvee_{i:x_i \neq z_i^1} \left(\kappa_i(x_i \Upsilon + \bar{x}_i \bar{\Upsilon}) = 2 \right) \vee \bigvee_{i:x_i = z_i^2} \left(\kappa_i(x_i \Upsilon + \bar{x}_i \bar{\Upsilon}) = -2 \right) \vee \bigvee_{i:z_i^1 \neq z_i^2} \left(\kappa_i(x_i \Upsilon + \bar{x}_i \bar{\Upsilon}) = -2\Theta \right).$$

Suppose the dominating attribute k is shared (so $\Theta = 1$). Then our second example implies $(\kappa_1 = 1) \vee (\kappa_2 = 1)$, an implicit preference favoring PhDs, or men. Our fourth example implies $\kappa_2 = 1$, an implicit preference favoring men.

Examples without direct comparisons. It is straightforward to identify a single implicit preference without ever observing a "direct" comparison (a comparison with only one non-shared attribute). See e.g. our third and fourth examples of parallel convex scissors.

Other Examples. Finally we give five examples that are either inconclusive about implicit preferences, or falsify our assumptions. An **equilateral triangle** and a **non-convex scissor** are examples of datasets in which Dilution does not rank any of the comparisons. Thus wins neither influence-dominate losses, nor vice versa. Corollary 2 tells us they cannot be rationalized by standard preferences, but without further assumptions we cannot rule out any other $\boldsymbol{\kappa}$. A **square** cycle, a pair of **opposing triangles**, and a pair of **opposing scissors** are examples of datasets in which each win can be matched to a loss with the same shared and non-shared attributes and hence, by Equivalence, the same influence. Thus wins influence-dominate losses and losses influence-dominate wins for all attributes. Corollary 3 tells us no $\boldsymbol{\kappa}$ can rationalize these datasets, given our assumptions on $\sqsupseteq_i$.

Equilateral	Square	Opposing triangles	Non-convex scissor	Opposing scissors
$\boldsymbol{\kappa} \neq 0$	Falsification	Falsification	$\boldsymbol{\kappa} \neq 0$	Falsification

4   Foundations

We now provide models of three types of decision maker that are consistent with our theory. We briefly summarize how each conforms to our core intuition. In the *ceteris paribus* foundation, an implicit preferences is a true preferences that is constrained by a rule, which applies when certain attributes are shared. Diluting the non-shared attributes can cause rules to

switch off, allowing the decision maker to express his true preferences. In the *signaling* foundation, an implicit preferences is a true preference that is concealed due to a signaling motive. When its attribute is diluted, the observer learns less about the decision maker's preference for that attribute, so he can more freely express his true preference. In the *implicit associations* foundation, an implicit preferences is an unconscious positive or negative association with an attribute, that the conscious brain would like to adjust for. The more an the attribute is diluted, the harder it is to distinguish between possible associations, so the less the decision maker can adjust for them.

To keep the discussion concise, for each foundation we provide the setup of the model and state the main result, that the foundation implies an Implicit Preferences utility function (i.e., consistent with (1)), with an influence function satisfying Equivalence and Dilution (Assumptions 1 and 2). At the end, we provide conditions under which each foundation also satisfies Assumption 3. Derivations and proofs are provided in web appendix A.2.

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.

4.1 *Ceteris Paribus* Decision Maker

Suppose our hiring manager freely chooses whichever candidate they prefer, except when comparing a man to an otherwise identical woman, in which case they are required to hire the woman. We state a general model of *ceteris paribus* decision makers who are constrained by rules that state they should favor certain attribute values "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 personal rule) or external (e.g. a bureaucratic rule).[17]

Multiple rules 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 supposed 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.

---

[17]For example, job advertisements at the Norwegian School of Economics (NHH) routinely include the sentence "In the event of equivalent qualifications, female applicants will be given preference."

Definition 11. *A **ceteris paribus utility function** has the form:*

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

for some $g : \{-1, 1\}^n \rightarrow \mathbb{R}$*, 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 higher scores when they are compared to otherwise-identical men.

Proposition 1. $u^{CP}(\boldsymbol{x}, \boldsymbol{z})$ *implies an Implicit Preferences utility function satisfying Equivalence and Dilution.*

Consider again the manager who prefers male candidates but is penalized for choosing a man over an otherwise identical women. Then, he will tend to favor men when gender is diluted (causing the rule to switch off), implying an *implicit* preference favoring men. Note that the implicit preference has the opposite sign to the penalty λ_i.

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 weights, and forms posteriors $\hat{w}_i$ based on the decision maker's actions. The core intuition is that the more other attributes share status with i, the less the observer learns about w_i, so the decision maker's signaling incentives weaken.

The observer's information differs between choice and evaluation, so we describe separate models for each. 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).

Signaling-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 *non-shared* attributes. We make two assumptions, which amount to the observer expecting the decision maker to be indifferent *ex ante*.[18] 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 (otherwise they would expect a particular bundle to be chosen). These are quite strong assumptions, but our conclusions should extend to small deviations. We discuss their relevance to applications in Section 5.

We define a 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 assume that $\boldsymbol{x}$ and $\boldsymbol{z}$ are distinct, so there is at least one non-shared attribute. We also assume that all preferences are expressed strictly.[19]

Definition 12. *A **signaling-choice utility function** has the form:*

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

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).

λ_i captures the decision maker's utility of shifting the observer's posterior over w_i. The separable setup implies the observer will only update about the weights on non-shared attributes.

Proposition 2. *$u^{SC}(\boldsymbol{x}, \boldsymbol{z})$ implies an Implicit Preferences utility function satisfying Equivalence and Dilution.*

Consider a 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. The implicit preference 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.

---

[18]If the observer had reason to believe the decision maker prefers one bundle over another, more dilute comparisons can sometimes be *more* informative about an attribute rather than less. For example, choosing a male PhD over a female MBA is plausibly 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 clearly *more* informative, in the sense of posteriors being farther apart.

[19]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 or 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 derive for the 2-action world correctly predicts behavior in a 3-action world, so the model can be applied to data containing indifferences.

Signaling-Evaluation. In evaluation we assume the decision maker reports their utility of two bundles, $\boldsymbol{x}$ and $\boldsymbol{z}$, 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 the observer has constant priors over the intrinsic values, nor that they are naïve.

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

Definition 13. *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 $g : \{-1, 1\}^n \to \mathbb{R}$, $\boldsymbol{\lambda} \in \mathbb{R}^n$, $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \dots, \sigma_n^2))$.

Lemma 1. *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:*

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 $U^2 = -E\left[\sum_{i=1}^{n}(\hat{w}_i - w_i)^2 \Big| y^x, y^z\right]$, with $g(\cdot)$ and $\boldsymbol{\lambda}$ common knowledge, and priors $\boldsymbol{w} \sim N(0, diag(\sigma_1^2, \dots, \sigma_n^2))$.

λ_j captures the decision maker's utility of shifting the observer's posteriors over w_j, while σ_j^2 is the variance of the observer's prior on w_j. The final term in u^{SE} captures how the decision maker adjusts her evaluations to influence the observer's beliefs. The adjustment to attribute i is proportional to the observer's uncertainty about w_i (σ_i^2), and inversely proportional to the total uncertainty about the weights on attributes with the same status as i.[20]

Proposition 3. *$u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ implies an Implicit Preferences utility function satisfying Equivalence and Dilution.*

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.

---

[20]Unlike Signaling-Choice, we solved the model assuming a sophisticated observer. The quadratic loss function means that player 1's optimal strategy is independent of the observer's priors on $\boldsymbol{\lambda}$, so our solution continues to hold if the observer has incorrect priors, including full naïveté (believing $\boldsymbol{\lambda} = \mathbf{0}$).

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. This model is based on Cunningham (2016). The true value of bundle $\boldsymbol{x}$ is given by:

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

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 reports expected values for $\boldsymbol{x}$ 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 intuition is that associations are generally informative (otherwise agent 2 would just ignore agent 1's estimates). However, agent 2 has access to contextual information that leads her to adjust agent 1's estimates. For example, let $x_i = 1$ for men. The decision maker might have an unconscious positive attitude toward men ($\lambda_i > 0$). However, the conscious brain believes that in the current setting, gender associations are normatively irrelevant ($\pi_i = 0$), so she would like to adjust agent 1's reports to account for this. Her ability to apply this adjustment depends on the degree to which she can separately distinguish the effect of λ_i from possible associations on other attributes.

Definition 14. *An **implicit associations utility function** has the form:*

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

with

$$\pi_i \in \mathbb{R}_+ \ \& \ E[\pi_i] = 1 \qquad \text{(1st agent's priors)}$$
$$\boldsymbol{\lambda} \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2)) \quad \text{(2nd agent's priors)}$$
$$\boldsymbol{\pi} \perp\!\!\!\perp \boldsymbol{\lambda} \qquad \text{(independence of priors)}.$$

u^{IA} represents the second agent's best guess at the true value $f(\boldsymbol{x})$.

In this model the sensitivity of utility to attribute i will be proportional to a weighted average of the adjustments (the πs) on all attributes with the same status as i. Thus a dilution of attribute i can either increase or decrease influence depending on whether the dilution increases or decreases that weighted average. This is inconsistent with Assumption 2. The foundation satisfies Assumption 2 in two special cases: (i) when at most one i has either $\lambda_i \neq 0$ or $\pi_i \neq 1$; or (ii) when there are exactly two attributes ($n = 2$). We adopt the first assumption for the remainder of the section, which implies that there can be an implicit preference for at most one attribute. For the second, a derivation is available on request.

Proposition 4. $u^{IA}(\boldsymbol{x}, \boldsymbol{z})$ *implies an Implicit Preferences utility function satisfying Equivalence and Dilution if at most one attribute has a non-zero implicit association and/or non-unitary adjustment factor:* $\sum_{i=1}^{n} \mathbf{1}\{(\lambda_i \neq 0) \vee (\pi_i \neq 1)\} \leq 1$.

For intuition of how implicit associations can be interpreted as implicit preferences consider the hiring manager that has a positive association with male candidates ($\lambda_i > 0$), but believes gender is normatively irrelevant ($\pi_i = 0$). When the candidates differ only on gender, the influence of implicit preferences is low, as the second agent can directly detect and override the influence of λ_i. As gender is diluted, the agent 1's high valuation of a male candidate could be explained by other possible associations, that might not be normatively irrelevant. Agent 2 therefore only partially adjusts agent 1's reports. Thus λ_i influences their decision, increasing the utility of the man and manifesting as an implicit preference favoring men. Note that there is only an implicit preference if both $\lambda_i \neq 0$ *and* $\pi_i \neq 1$: the first agent must have a nonzero implicit association and the second agent must want to adjust it.[21]

4.4 Dominance of Attribute k

Finally, we give sufficient conditions for Assumption 3 to hold in all foundations that apply to evaluation (recall that the assumption is not relevant to choice).

Proposition 5. *The* Ceteris Paribus *decision maker of Proposition 1 satisfies Assumption 3 if k is not governed by a rule. The* Signaling-Evaluation *decision maker of Proposition 3 satisfies Assumption 3 if $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$. The* Implicit Associations *decision maker of Proposition 4 satisfies Assumption 3 if $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$.*

---

[21]The sign of the implied implicit preference depends on $\lambda_i(1 - \pi_i)$. If $\pi_i > 1$, the second agent wants to *amplify* their implicit associations (they think the first agent is too conservative). This generates an implicit preference with the opposite sign to λ_i. In our example, it would increase the value of men when the candidates differ only on gender, weakening as gender is mixed with other attributes.

5 Guidance for applications

We now discuss some practical guidance for applications of our theory.

Multivalued Attributes. Some attributes might take multiple values. For example, we might observe job candidates with three different qualifications (MBA/PhD/JD) instead of two. The data then need to be transformed to apply our theory. The appropriate transformation depends on the setting. Attribute values could be grouped together, or the dataset partitioned to focus on parts of the attribute space. In our analysis of Exley (2016)'s data, there are multiple lotteries with different win probabilities. We construct binary attribute spaces around each probability, and analyze them separately.

Ambivalence **in Choice Data.** In choice data a consideration arises that we call *ambivalence*. We recommend focusing on choice sets where participants are *likely to be close to indifferent*, for two reasons. First, identification relies on observing intransitive choices, which is unlikely when there is large variation in explicit values $v(.)$. Second, our signaling-choice foundation assumes the observer has equal priors over the utility of both bundles.

When there are multiple "non-ambivalent" attributes in the dataset we can group them together so that their combination plausibly satisfies ambivalence. For instance, while a hiring manager is unlikely to be close to 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 in Section 6.1 faces this issue. The basic attributes that vary in her experiment (e.g. higher and lower monetary amounts) are unlikely to satisfy ambivalence, so we group them together in such a way as to restore it.

Within-subjects Data. The theory assumes we observe the revealed preferences of a single decision-maker, that is, we observe within-subjects data. A concern in such datasets 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 systematic within-subject inconsistencies.[22]

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

Between-subjects Choice Data. Establishing the presence of intransitivities in between-

---

[22]A related concern is experimenter demand effects: participants may guess what the experimenter is looking for from the sequence of decisions they observe. Recent work that directly manipulates such beliefs finds mostly modest effects (de Quidt et al., 2018; Mummolo and Peterson, 2018).

subjects choice data is challenging, because intransitivity is difficult to distinguish from underlying heterogeneity in preferences (similar to the Condorcet paradox in pairwise voting). One (strong) remedy is to assume homogeneous preferences. Alternatively, one can test for violations of the Triangle inequality (see Regenwetter et al. (2011)). If each participant makes one choice, 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 $3/4$. It may be hard to find a setting with sufficiently strong intransitive preferences to satisfy such conditions (Müller-Trede et al., 2015).[23]

Between-subjects Evaluation Data. Our tools for evaluation data carry over well to between-subjects data with some functional form restrictions. 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 only by the structure of the comparison.[24] We also assume evaluations are affine in utility: $y(\boldsymbol{x}, \boldsymbol{z}) = a + bu(\boldsymbol{x}, \boldsymbol{z})$. Normalizing $a = 0, b = 1$, the mean evaluation is:

$$\frac{1}{T} \sum_{t=1}^{T} \left[v_t(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \kappa_{i,t} \theta_i(\boldsymbol{x}, \boldsymbol{z}) \right] \xrightarrow[T \to \infty]{} \overline{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i sgn(\overline{\kappa}_i) |\overline{\kappa}_i| \theta_i(\boldsymbol{x}, \boldsymbol{z}). \qquad (2)$$

This is equivalent to the utility function of a representative agent with implicit preferences $\kappa_i^{rep} = sgn(\overline{\kappa}_i)$ and influence function $\theta_i^{rep} = |\overline{\kappa}_i| \theta_i$. Thus our usual tools can identify $sgn(\overline{\kappa}_i)$, the sign of the *average* implicit preference in the population. If $sgn(\overline{\kappa}_i)$ is positive, we learn that at least some part of the population has positive κ_i. If we also assume that implicit preferences are aligned in the population (have weakly the same sign), we learn that sign.

6 Applications

6.1 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 choose between lotteries and sure payments, where the beneficiaries can be either themselves, or charity.[25] She uses those choices to construct certainty equivalents,

---

[23] 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.

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

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

such that each lottery to self, and each lottery to charity, is valued both in terms of money to self and in terms of money to charity. She then tests for variation in these certainty equivalents as the trade-off between self and charity varies. The dominant pattern is one in which participants tolerate more risk when the risk favors them (high certainty equivalents), and tolerate 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, suggesting *implicit selfishness*.

Reanalyzing Exley's dataset, we confirm this interpretation: 51 percent of participants reveal an implicit selfish preference. Our approach also yields new insights in the form of *implicit risk preferences*. 30 percent of participants become more risk averse when risk is diluted, while 15 percent become more risk tolerant. An important difference between our approaches is that Exley analyses differences in average lottery valuations across comparisons, which can give a measure of how strongly implicit preferences influence behavior, whereas we look for individual-level inconsistencies.[26] That allows us to more precisely classify individuals, and distinguish heterogeneity from average behavior (e.g., distinguishing implicitly risk averse and risk tolerant individuals).

Data. We need to do a little work to place Exley's data in a binary attribute framework. Web appendix A.3 provides a full description of the data structure, how we represent it using binary attributes, and how we use the same assumptions as Exley's analysis to impute certain choices that are not directly observed. Here we provide a brief summary.

Exley's dataset consists of an initial *normalization* choice to roughly calibrate the participant's exchange rate between money to themselves ("self") and money to charity. It elicits an amount $X payable to charity that is slightly preferred to $10 payable to self.

Each subsequent choice is between a safe payoff and a lottery paying a prize with probability P. Participants make four kinds of choices: (1) charity gets safe vs charity gets lottery; (2) self gets safe vs self gets lottery; (3) charity gets safe vs self gets lottery; (4) self gets safe vs charity gets lottery. For each question the participant chooses the highest safe amount that they would accept (using a "choice list"). All four questions are repeated for seven values of P: $\{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. We do not observe choices between pairs of lotteries with different probabilities, so we perform separate analysis for each value of P. Thus, each participant has seven separate opportunities to reveal their implicit preferences.

For each P, we can represent the data as four binary choices over two binary attributes, which we label Risk $\in$ {Safe, Risky}, and Social $\in$ {Generous, Selfish}.[27] The observed

---

[26]Exley does some individual-level analysis and reports "for 65% of the participants, the average difference in their lottery valuations across contexts is in a self-serving direction." That does not guarantee they exhibit the "parallel triangles" necessary to identify implicit Selfishness in our setup.

[27]Bundles differ in prize, probability of winning, and recipient, which by themselves do not satisfy Am-

choice sets are shown in black in the diagrams below. Not every choice set on this binary space is observed in the data. Specifically, participants do not make choices between bundles that only differ in Risk (the horizontal choice sets in the diagrams below). We need these to detect implicit social preferences.[28] Exley's analysis faces a similar issue as she needs to compare lottery valuations elicited in dollars to self to those elicited in dollars to charity. She uses the participant's value of X, plus a *linearity in payoffs* assumption, to do this. That same assumption allows us to impute the choice (Generous, Safe) $\succ$ (Selfish, Safe), marked in blue on the diagrams below (see Appendix). Because the choice lists are calibrated from participants' individual values of X, tilting payoffs to slightly favor Generous, we do not observe the choices we would need to impute the opposing preferences.

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.

Classifying Individuals by Implicit Preference Type. Given the choice sets we observe, there are three patterns of choice that can unambiguously identify an implicit preference. Panels (a) and (b) below show figure 8 cycles that reveal implicit preferences on the Risk attribute. Panel (c) shows a pair of parallel right triangles that reveals an implicit preference for Selfish. As noted above, we do not observe the opposing horizontal choice so cannot detect implicit Generous preferences.

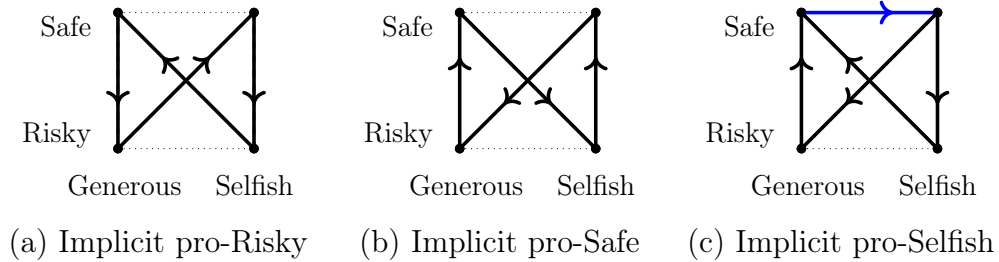(a) Implicit pro-Risky   (b) Implicit pro-Safe   (c) Implicit pro-Selfish

Table 1 presents the empirical frequencies of each type of cycle, averaged over the 86 participants and 7 values of P (602 observations). Overall, participants exhibit one of the cycles of interest 33 percent of time, but at different frequencies. Only 5 percent of choices exhibit pro-Risky cycles, 10 percent are pro-Safe, while 18 percent are pro-Selfish.

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

---

bivalence (see Section 5). We construct binary attributes such that Safe bundles pay small prizes for sure, while Risky bundles pay larger prizes with probability P. Selfish bundles pay small prizes to self, while Generous bundles pay larger prizes to charity. See the Appendix for full details.

[28]Any pattern of choice in the choice sets that we observe can be consistent with entirely *explicit* selfish preferences. E.g., those in panel (c) can be rationalized by (Selfish, Safe) $\succ$ (Selfish, Risky) $\succ$ (Generous, Risky) $\succ$ (Generous, Safe), which always ranks Selfish above Generous.

Table 1: Frequencies of different cycles in Exley (2016) data

Cycle			Frequency	s.e.	95% CI
Implicit pro-Risky	(a)		.048	(.01)	[.032, .073]
Implicit pro-Safe	(b)		.103	(.018)	[.073, .144]
Implicit pro-Selfish	(c)		.181	(.024)	[.139, .233]

This table shows the frequency of each type of the cycle in our analysis of Exley (2016). Standard errors clustered at the participant level. Statistical tests: $p(a = b) = .009$, $p(a = c) < .001$, $p(b = c) = .019$, $p(a = b = c) < .001$. Test against random choice benchmark: $p(a + b + c = .25) < .001$.

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 2a plots 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.

(a) Implicit risk preferences (b) Implicit social preferences
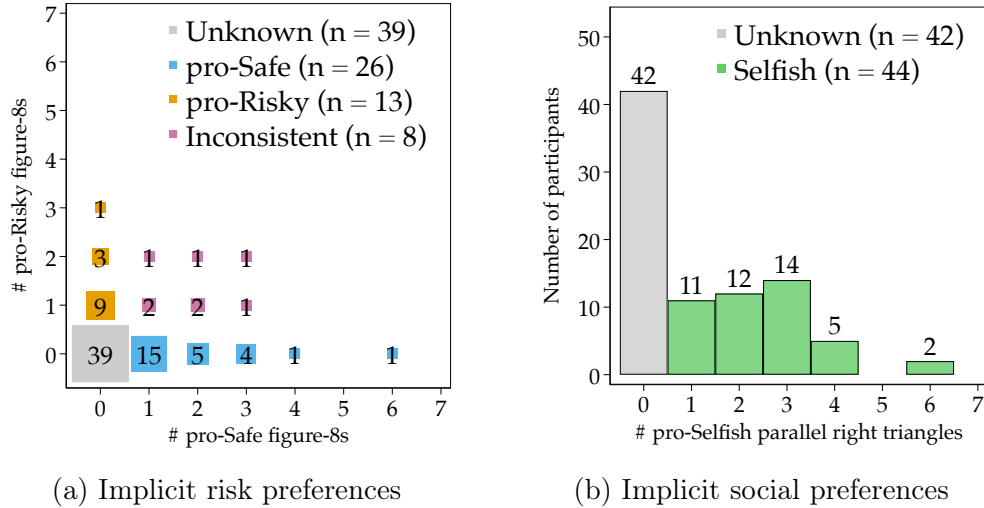
Panel (a) classifies participants according to their number of implicit pro-Risky and pro-Safe cycles. Cell size and numeric labels indicate the number of participants in each cell, colors indicate the type classification. Panel (b) classifies participants according to their number of implicit pro-Selfish cycles (the data structure means that we do not observe pro-Generous cycles).

Figure 2: Type classifications in Exley (2016) data

Implicit risk attitudes are prevalent in the sample, and tend to be implicitly risk-averse. This could have important implications for real-world decisions. For example, an implicitly risk-averse decision-maker might make more risk-averse choices when choosing between pension

plans with different attributes (where implicit risk preferences have high influence) than she would when choosing between different variants of the same plan (where influence is lower). That could have substantial implications for wealth at retirement.

However, relatively few participants exhibit more than one of these cycles. Of the consistent participants, 11 exhibit two or more pro-Safe cycles, while 4 exhibit two or more pro-Risky cycles. This suggests implicit risk attitudes may be weak at the individual level.

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 2b shows that 51% of participants (44 in total) are classified as implicitly pro-Selfish, of whom 33 exhibit two or more pro-Selfish cycles. Implicit selfishness is more widespread, and expressed more frequently, than either pro-Safe or pro-Risky implicit preferences. However, we cannot assess the extent of inconsistency in this preference.

Statistical Analysis. Our analysis so far assumes behavior is deterministic, but in reality some of the heterogeneity we observe is likely a result of errors, or noise in the data. We first verify that the data are inconsistent with purely random behavior. Table 1 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 pair of frequencies. We observe a strong systematic tendency toward pro-Selfish cycles, and a strong tendency toward pro-Safe relative to pro-Risky cycles.[29] In Web Appendix A.3.5 we perform permutation tests to examine whether the data are consistent with homogeneous implicit preferences plus noise. We conclude that there is significant evidence of both systematic and heterogeneous implicit preferences in the sample.

Web Appendix A.3.5 reports two more analyses. First, we ask whether implicit social and risk preferences are correlated. They appear not to be: we find no meaningful difference in the distribution of implicit risk preferences between those classified as implicitly selfish or not. Second, Exley collected a small number of descriptive variables; we regress them on our type classifications. We find (similar to Exley) that implicitly selfish participants are significantly more likely to exploit "moral wiggle room," in a task modeled on Dana et al. (2007). Interestingly, so are implicitly risk averse participants, perhaps because the moral wiggle room task directly leverages uncertainty about the consequences of one's actions.

In sum, like Exley, we find substantial evidence of implicit selfishness. Our analysis demonstrates the applicability of our method to experimental choice data, and how it can

---

[29]We also simulate a large dataset where the switching point in each choice list is uniformly random. In the simulated dataset, each type of cycle occurs with frequency .083 (well outside the 95% CIs for type (a) and (c) cycles), and the frequency of at least one cycle is .25 (in the data the rate is 0.33, with 95% CI [.278, .386]). Thus the observed behavior differs substantially from this random choice benchmark.

extract new findings (implicit risk preferences) from data collected for another purpose.

6.2 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 reanalyze DeSante's data using our framework, to test for implicit racial preferences. Specifically, we will test whether participants tend to award more money to applicants of one race, and less to the other, when the influence of implicit racial preferences increases.

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.[30]

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).[31] Some participants evaluate two applicants of the same race, while others evaluate one from each race. Second, in some conditions there is also an assessment of each applicant's Work Ethic $\in$ {Good, Bad}.[32] When reported, this attribute is always non-shared. 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 Children $\in$ {c, c'} which always differs within the comparison set, with no implicit preference attached to it.

Below we represent the data structure graphically. We observe some evaluations of 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, c) is evaluated alongside (Black, Good, c') and (White, Good, c'). From these we can construct six convex scissors. We draw the inequalities that would be implied by an implicit preference favoring White applicants.

---

[30]The 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. We expect this to make it harder to detect implicit preferences as it particularly constrains 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.

[31]Simonsohn (2016) points out that names might signal something additional to race, e.g. socioeconomic status. So we might be observing implicit preferences over SES instead of race.

[32]The language in the experiment is "Excellent/Poor", we use "Good/Bad" for compactness.

(a) Work ethic concealed   (b) Work ethic revealed
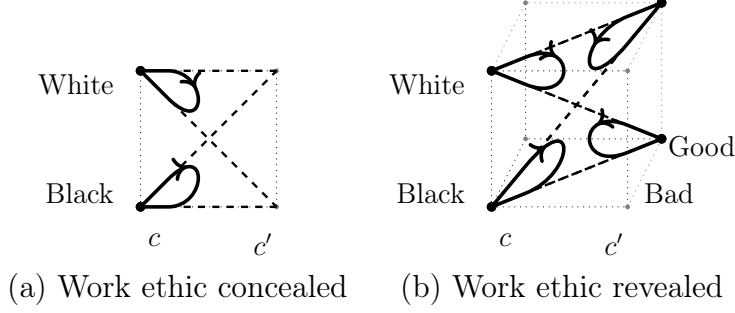
Dilution (Assumption 2) does not rank the influence of implicit racial preferences between comparisons within a scissor, because race switches status from shared to non-shared. Instead, we assume there is an attribute satisfying Assumption 3 that is shared in all comparisons (e.g., all applicants are women). Then, the influence of racial preferences is higher when race is shared than non-shared. This is intuitive: many other attributes could explain why evaluations are generally low or generally high when race is shared.

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 both cases evaluations are expected to 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 the diagrams above. We therefore cannot identify implicit preferences at the individual level. Instead we compare mean evaluations between comparisons. Imposing linearity, we can interpret these as revealing mean implicit preferences in the sample (see Section 5).

Figure 3 presents the results. We group evaluations in pairs that correspond to the six convex scissors. Under concealed work ethic the two scissors constitute a pair of parallel convex scissors. Under revealed work ethic there are several ways to construct parallel convex scissors. We examine them collectively.

We find positive differences in five out of six scissors (colored in blue), meaning that the general pattern is consistent with an implicit preference favoring White applicants. Only one is statistically significant but the average difference equals \$34 and is highly significant ($p < 0.01$). An F-test of the null that all six differences equal zero has a p-value of 0.08.

We can go further and estimate some parameters of interest directly (see web appendix A.4 for details). The influence of implicit racial preferences is highest when race is shared, while influence of work ethic is highest when race is non-shared. Denote higher influence values by $\theta_{race}^H, \theta_{ethic}^H$ and lower ones by $\theta_{race}^L, \theta_{ethic}^L$.[33] Exploiting linearity (equation (2)) we

---

[33]Note that the values of θ_{race}^H and θ_{race}^L will also differ between concealed/revealed work ethic treatments.
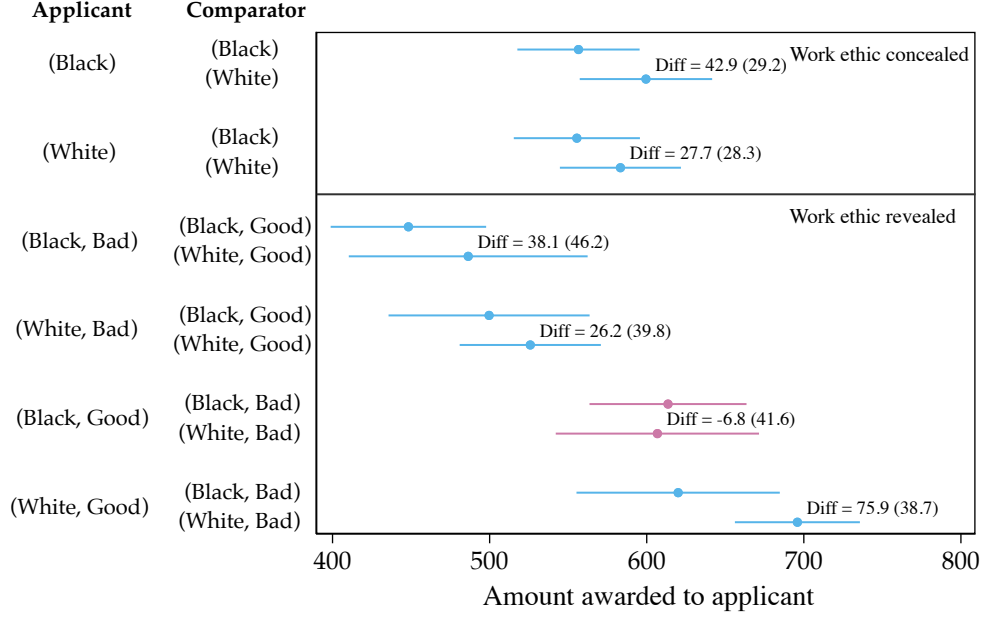
Each pair of points corresponds to one "convex scissor." pro-White implicit preferences imply positive "Diffs" (shown in blue). 95% confidence intervals clustered by participant. $N = 753$ participants, 1,506 evaluations.

Figure 3: Reanalysis of DeSante (2013) data

can identify $2 \times \overline{\kappa}_i \left(\theta_i^H - \theta_i^L \right)$, which corresponds to the mean increase in evaluation of a bundle with $x_i = 1$ relative to a bundle with $x_i = -1$, when influence increases from θ_i^L to θ_i^H. In words, it measures the widening of the gap between Black versus White (or Good versus Bad), when influence increases, which we interpret as driven by implicit preferences.

Table 2 presents our findings. When race switches from non-shared to shared the gap between White and Black candidates increases by \$71 ($p = .02$) in the Concealed work ethic treatment, and by \$66 ($p = .08$) in the Revealed work ethic treatment. We also find that increasing influence of work ethic decreases the gap between Good and Bad candidates, by \$47, but this is not significant ($p = .18$).[34] In conclusion, we find significant evidence of implicit pro-White preferences, and modest evidence of implicit preferences over work ethic.

7   Related Theories

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

---

[34]Three out of the four convex scissors (the first, third, and fourth) are directionally consistent with an implicit preference favoring Bad candidates (i.e. Bad candidates are penalized less when influence is high).

Table 2: Quantitative estimates using Desante (2013) data

	Work ethic concealed	Work ethic revealed
$2 \times \overline{\kappa}_{race} \left(\theta_{race}^H - \theta_{race}^L \right)$	71.13**	66.16*
	(30.21)	(38.10)
$2 \times \overline{\kappa}_{ethic} \left(\theta_{ethic}^H - \theta_{ethic}^L \right)$		-47.27
		(35.33)
Participants	378	375

$2 \times \overline{\kappa}_{race} \left(\theta_{race}^H - \theta_{race}^L \right)$ equals the increase in evaluation of White candidates, relative to Blacks, when influence increases from θ_i^L to θ_i^H. Second row measures the same for Good relative to Bad candidates. Standard errors clustered by participant in parentheses, $^*p < .1, ^{**}p < .05$.

Contingent weighting. Models of contingent weighting in multi-attribute choice allow preferences to depend on the choice set, in common with our theory. 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). However in most such theories the sensitivity with respect to each attribute depends on the marginal distribution of realizations of that attribute in the choice set, while in our model it depends on the *joint* distribution across all attributes. Concretely, we are not aware of a contingent-weighting model that can generate a figure-8 intransitivity, which we think epitomizes the conflict between conflicting underlying preferences that motivates our theory.[35]

Complexity/inattention. Our identification comes from comparing decisions where more or fewer attributes vary. If the former are more complex than the latter, we might 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. 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

---

[35] 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 our leading 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. Separability by attribute holds for each of the models discussed above except Bordalo et al. (2013), but to the best of our knowledge that model is not consistent with intransitive cycles in binary choices with two attributes (Ellis and Masatlioglu (2021)).

cycles over non-monetary attributes.[36]

Evaluability. A similar point applies to the literature 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."[37] Again, this is a quite different principle to that used in this paper. Increased/decreased sensitivity to an attribute could not generate a figure-8 cycle, by an analogous argument to footnote 35.

We can model separate evaluation in our framework as the evaluation of $\boldsymbol{x}$ with only itself as comparator: $y(\boldsymbol{x}, \boldsymbol{x})$. Then, all attributes are shared ($|\boldsymbol{x} - \boldsymbol{x}| = \boldsymbol{0}$). In all of our foundations this maximizes the influence of implicit preferences, which could lead to decreased sensitivity to attributes where implicit and explicit preferences oppose one another.

Inference from the choice set. We assume that the attribute values of one bundle are uninformative about the value of other bundles. If not then 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. Take our leading example: the manager's decisions could be rationalized if they (1) prefer women to men; but (2) believe that qualifications are more valuable if they are typically male. Thus in the diagonal choice sets they prefer the man not because he is a man but because he has the qualification that men have. The explanation seems a stretch: it 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). Moreover, in applications with familiar attributes the scope for learning from the choice set seems small.

8 Conclusion

Our paper is motivated by an assumption that is latent in a number of empirical papers: that people sometimes hold two opposite preferences regarding an attribute and that one preference—the *implicit* preference—has greater influence when the comparison mixes that attribute with others. By formalizing the assumption we are able to give precise guidance

---

[36]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.

[37]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 developed in Hsee and Zhang (2010). See Cunningham (2013) for a Bayesian rationalization of increased sensitivity in joint evaluation.

for inferring the direction of a decision maker's implicit preferences from their decisions, in a way that is applicable to many existing empirical datasets.

Some possible extensions to our framework include generalizing the representation theorem to weaken the separability assumptions, investigating alternative influence-dominance relations, extending the attribute space to nonbinary attributes, and allowing for bundles that are missing some attributes. For instance, we could add a "no hire" bundle to the hiring setting. We conjecture that the influence of implicit gender preferences is stronger in a choice between "hire a woman" and "no hire" than between "hire a woman" and "hire a man." Then, a manager that implicitly favors men might exhibit the cycle $Female \succ Male \succ No\ hire \succ Female$.

It is natural to ask how implicit preferences will be revealed in comparison sets larger than two elements. In attempting to answer these questions we find that the predictions of the different foundational models diverge. For example, in the context of choice we find that influence is most naturally interpreted as a property of a *choice set* in the implicit associations model, because the more estimates are reported by the first agent, the more the second agent can infer about her associations. But it is most naturally interpreted as a property of the *choice* in the signaling and ceteris-paribus models, because the observer or rule-setter's information comes from what is chosen and what is not. When considering binary choice sets the distinction does not matter, because when one bundle is rejected the other is always chosen. Thus our *general* concept of an implicit preference, which is agnostic about the underlying foundation, is limited to binary comparisons.

Our utility function is compatible with other theories about what affects implicit preferences' influence (θ_i), e.g. variation in time pressure, stated versus revealed preferences, or moral wiggle room. A central advantage of our definition is that θ_i depends *only* on the comparison $(\boldsymbol{x}, \boldsymbol{z})$; we not require any information beyond what is contained in the bundles themselves, widening applicability and reducing degrees of freedom. To the extent that other factors also predict variation in influence, we would expect the implicit preferences identified using different assumptions to coincide.

We see rich scope for empirical applications, through data reanalysis as well as fresh experiments, to systematically map out the existence, strength of influence, consistency, and out-of-sample predictiveness of implicit preferences across many diverse domains. Our applications found evidence of implicit selfishness, implicit risk preferences, and implicit racial bias. Figure 1 suggests some additional domains that we see as promising, including temptation, embarrassing decisions, prejudice (in many settings), framing, and time discounting. We particularly highlight the Framing example. There, we conceptualize a *frame* as an attribute over which the decision maker has zero explicit preference but a nonzero implicit

preference. Thus they are indifferent between identical but differently-framed prospects, but frames influence choice when mixed with other attributes. These, and other applications, we leave to future research.

9 Appendix: Proof of Theorem 1

To prove the theorem we will need to define several vectors and matrices that are indexed by comparisons $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{X}$. This can quickly make the notation unreadable, so we introduce a shorthand that we use whenever it does not generate ambiguity. We use $\delta = (\boldsymbol{x}, \boldsymbol{z})$ to represent a generic comparison $(\boldsymbol{x}, \boldsymbol{z})$, allowing us to write, e.g., $M_{i,(\boldsymbol{x},\boldsymbol{z}),(\boldsymbol{x}',\boldsymbol{z}')}$ as $M_{i,\delta,\delta'}$. Similarly, if $(\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')$ we write $\delta \sqsupseteq_i \delta'$. The set of all δs is the same as the set of all comparisons: $\delta \in \mathcal{X} \times \mathcal{X}$, which has $|\mathcal{X}|^2$ elements.

Each inequality in dataset D can written as:

$$v(\boldsymbol{x}^j) + \sum \boldsymbol{x}_i^j \kappa_i \theta_i(\boldsymbol{x}^j, \boldsymbol{z}^j) \geq v(\boldsymbol{x}'^j) + \sum \boldsymbol{x}_i'^j \kappa_i \theta_i(\boldsymbol{x}'^j, \boldsymbol{z}'^j),$$

where the inequality is strict for $j \leq \bar{m}$. We can write the two functions, $v(\cdot)$ and $\theta_i(\cdot)$ as vectors $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{X}|}$ and $\boldsymbol{\theta} \in \mathbb{R}^{n|\mathcal{X}|^2}$, with elements $v_x = v(\boldsymbol{x})$ (one entry for each $\boldsymbol{x} \in \mathcal{X}$), and $\theta_{i\delta} = \theta_i(\delta)$ (one entry for each $i \in \{1, ..., n\}$ and comparison $\delta \in \mathcal{X} \times \mathcal{X}$).

We can now state the problem as follows. The vector of implicit preferences $\boldsymbol{\kappa}$ rationalizes D if and only if there exist vectors $\boldsymbol{v}$ and $\boldsymbol{\theta}$ such that (1) every inequality in D is satisfied, and (2) $\boldsymbol{\theta}$ obeys influence-dominance, meaning $\delta \sqsupseteq_i \delta' \implies \theta_{i\delta} \geq \theta_{i\delta'}$.

We can write D's inequalities in matrix form with $\begin{bmatrix} \hat{P} & \hat{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{bmatrix} \gg 0$ representing the $\bar{m}$ strict inequalities, and $\begin{bmatrix} \bar{P} & \bar{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{bmatrix} \geq 0$ representing the $m - \bar{m}$ weak inequalities. Each row corresponds to one inequality. The matrix $P = \begin{bmatrix} \hat{P} \\ \bar{P} \end{bmatrix} \in \mathbb{Z}^{m \times |\mathcal{X}|}$ holds the coefficients on $\boldsymbol{v}$, with entries:

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

The matrix $X = \begin{bmatrix} \hat{X} \\ \bar{X} \end{bmatrix} \in \mathbb{Z}^{m \times n|\mathcal{X}|^2}$ holds the coefficients on $\boldsymbol{\theta}$, with entries:

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

Finally we encode the influence-dominance relations $\sqsupseteq_i, i = \{1, ..., n\}$ as a matrix of coefficients on $\boldsymbol{\theta}$: $Q \in \mathbb{Z}^{n|\mathcal{X}|^4 \times n|\mathcal{X}|^2}$. Q has one row for each combination of an attribute k and pair of comparisons $\bar{\delta}, \bar{\delta}'$ ($n|\mathcal{X}|^4$ rows in total). A row has non-zero entries only if $\bar{\delta} \sqsupseteq_k \bar{\delta}'$. If so, the row has entry $+1$ in the column that corresponds to attribute k and comparison $\bar{\delta}$, and

40

-1 in the column corresponding to attribute k and comparison $\bar{\delta}'$:

$$Q_{\underbrace{k\bar{\delta}\bar{\delta}'}_{\substack{\text{row}\\k\in\{1,\ldots,n\}\\\bar{\delta},\bar{\delta}'\in\mathcal{X}\times\mathcal{X}}},\,\underbrace{i\delta}_{\substack{\text{column}\\i\in\{1,\ldots,n\}\\\delta\in\mathcal{X}\times\mathcal{X}}}} = \mathbb{1}\Big\{\underbrace{(i=k)}_{\substack{\text{column}\\\text{corresponds to }k}} \wedge \underbrace{(\bar{\delta}\sqsupseteq_i\bar{\delta}')}_{\substack{\bar{\delta}\text{ influence-}\\\text{dominates }\bar{\delta}'}}\Big\} \times \Big(\underbrace{\mathbb{1}\{\delta=\bar{\delta}\}}_{\substack{=1\text{ if column}\\\text{corresponds to }\bar{\delta}}} - \underbrace{\mathbb{1}\{\delta=\bar{\delta}'\}}_{\substack{=1\text{ if column}\\\text{corresponds to }\bar{\delta}'}}\Big).$$

Then, the vector $\boldsymbol{\theta}$ obeys influence-dominance if and only if $Q\boldsymbol{\theta}\geq 0$. Putting the pieces together, we can say that $\boldsymbol{\kappa}$ rationalizes D if and only if the following Condition holds:

Condition 1. *There exists a real-valued vector $\begin{bmatrix}\boldsymbol{v}\\\boldsymbol{\theta}\end{bmatrix}$ satisfying*

$$\begin{bmatrix}\hat{P} & \hat{X}\end{bmatrix}\begin{bmatrix}\boldsymbol{v}\\\boldsymbol{\theta}\end{bmatrix} \gg \boldsymbol{0} \quad \text{(all positive)}$$

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

Motzkin's Rational Transposition Theorem (Border (2013)) tells us that Condition 1 will be **true** if and only if our next condition, Condition 2, is **false**. Condition 2 expresses that a non-negative-weighted sum of rows of $\begin{bmatrix}P & X\\0 & Q\end{bmatrix}$ yields a vector of zeroes.

Condition 2. *There exist integer-valued vectors $\hat{\boldsymbol{p}}\in\mathbb{Z}^{\bar{m}}$, $\bar{\boldsymbol{p}}\in\mathbb{Z}^{m-\bar{m}}$, $\boldsymbol{q}\in\mathbb{Z}^{n|\mathcal{X}|^4}$ (with $\boldsymbol{p}\equiv\begin{bmatrix}\hat{\boldsymbol{p}}\\\bar{\boldsymbol{p}}\end{bmatrix}$), satisfying:*

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

$$\hat{\boldsymbol{p}} > 0 \qquad \text{(all non-negative, at least one positive)}$$
$$\bar{\boldsymbol{p}} \geq 0, \boldsymbol{q} \geq 0 \quad \text{(all non-negative)}$$

Loosely speaking, given implicit preferences $\boldsymbol{\kappa}$, there exist vectors $\boldsymbol{v}$ and $\boldsymbol{\theta}$ that can rationalize the dataset if and only if there is no combination of rows in $\begin{bmatrix}P & X\\\boldsymbol{0} & Q\end{bmatrix}$), which exactly cancel.

We now prove 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 $\kappa_i=1$, losses influence-dominate wins, and for every $\kappa_i=-1$, wins influence-dominate losses.*

Proof that condition 3 implies condition 2. We construct vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ from the cyclical selection $\boldsymbol{s}$ and matching matrices $M_i, i=\{1,\ldots,n\}$, to show $\begin{bmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{bmatrix}\begin{bmatrix}P & X\\\boldsymbol{0} & Q\end{bmatrix}=\boldsymbol{0}^T$.

Let:
$$\forall j\in\{1,\ldots,m\}, \qquad p_j = s_j$$
$$\forall i\in\{1,\ldots,n\},\, \delta,\delta'\in\mathcal{X}\times\mathcal{X}, \quad q_{i\delta\delta'} = M_{i,\delta,\delta'}.$$

41

By the definition of a cyclical selection, $\hat{\boldsymbol{p}} > 0$ and $\bar{\boldsymbol{p}} \geq 0$, and by the definition of a matching, $\boldsymbol{q} \geq 0$. For each element of the vector $\boldsymbol{p}^T P \in \mathbb{Z}^{|\mathcal{X}|}$, which is indexed by $\boldsymbol{x}$, we can write:

$$\sum_{j=1}^{m} p_j P_{j,\boldsymbol{x}} = \sum_{j=1}^{m} s_j P_{j,\boldsymbol{x}} = \sum_{j=1}^{m} s_j \left(\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} - \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\} \right) = 0.$$

Where the first equality follows from the definition of $\boldsymbol{p}$, the second from the definition of P, and the third from the definition of a cyclical selection: each bundle $\boldsymbol{x}$ must appear equally often on the left- and right-hand side. Thus $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} P \\ \mathbf{0} \end{bmatrix} = \mathbf{0}^T$.

An element of the vector $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X \\ Q \end{bmatrix} \in \mathbb{Z}^{n|\mathcal{X}|^2}$, indexed by $(i\delta)$, can be expressed as:

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

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

$$\underbrace{\sum_{j:(x^j,z^j)=\delta} p_j x_i^j \kappa_i}_{\substack{\text{inequalities with} \\ \delta \text{ on LHS}}} - \underbrace{\sum_{j:(x'^j,z'^j)=\delta} p_j x_i'^j \kappa_i}_{\substack{\text{inequalities with} \\ \delta \text{ on RHS}}} + \underbrace{\sum_{\bar{\delta}':\delta \sqsupseteq_i \bar{\delta}'} q_{i\delta\bar{\delta}'}}_{\substack{Q \text{ rows where } \delta \\ \text{influence-dominates}}} - \underbrace{\sum_{\bar{\delta}:\bar{\delta} \sqsupseteq_i \delta} q_{i\bar{\delta}\delta}}_{\substack{Q \text{ rows where } \delta \\ \text{influence-dominated}}} \qquad (3)$$

Given $\boldsymbol{p} = \boldsymbol{s}$ the first two terms equal κ_i multiplied by the score for that i, δ pair:

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

Take the last two terms of (3) and substitute $q_{i\delta\delta'} = M_{i,\delta,\delta'}$. We obtain:

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

$$= -\kappa_i c_{i,\delta},$$

which uses Definition 5: the first equality follows from "matches obey dominance" and the second from "all scores are accounted for." Substituting into equation (3) we obtain $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} P & X \\ 0 & Q \end{bmatrix} = \mathbf{0}^T$, establishing Condition 2.

Proof that condition 2 implies condition 3. We construct a vector $\boldsymbol{s}$ and matrices $M_i, i = \{1, \ldots, n\}$ from the vectors $\hat{\boldsymbol{p}}, \bar{\boldsymbol{p}}, \boldsymbol{q}$, and show that they satisfy Definitions 3 and 5:

$$\forall j \in \{1, \ldots, m\}, \qquad s_j = p_j$$
$$\forall i \in \{1, \ldots, n\}, \; \delta, \delta' \in \mathcal{X} \times \mathcal{X}, \quad M_{i, \delta, \delta'} = q_{i\delta\delta'} \mathbb{1}\{\delta \sqsupseteq_i \delta'\}$$

We can verify that $s_j > 0$ for at least one $j \leq \bar{m}$ because $\hat{\boldsymbol{p}} > 0$, and that $s_j \geq 0$ and $M_{i, \delta, \delta'} \geq 0$ because $\bar{\boldsymbol{p}}, \boldsymbol{q} \geq 0$. To confirm that $\boldsymbol{s}$ is a cyclical selection we need to show that $\sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} = \sum_{j=1}^{m} s_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\}$. This follows because $\boldsymbol{p}^T P = \boldsymbol{0}^T$ (by condition 2), with elements (indexed by $\boldsymbol{x}$):

$$\sum_{j=1}^{m} p_j P_{j, \boldsymbol{x}} = \sum_{j=1}^{m} p_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^j\} - \sum_{j=1}^{m} p_j \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}'^j\},$$

where the equality comes from the definition of P. We must finally verify that for each i with $\kappa_i = 1$, losses influence-dominate wins, and for each i with $\kappa_i = -1$, wins influence-dominate losses. I.e., we check that M_i satisfies the conditions of Definition 5. Observe that:

1. Matches obey dominance: $\forall \delta, \delta' \in \mathcal{X} \times \mathcal{X}, \; (M_{i, \delta, \delta'} > 0) \implies (\delta \sqsupseteq_i \delta')$. This immediately follows because we constructed M_i from $\boldsymbol{q}$ as $M_{i, \delta, \delta'} = q_{i\delta\delta'} \mathbb{1}\{\delta \sqsupseteq_i \delta'\}$.

2. All scores are accounted for, i.e. for every $\delta \in \mathcal{X} \times \mathcal{X}$ and $i \in \{1, \ldots, n\}$ with $\kappa_i = 1$:

$$\sum_{\bar{\delta}' \in \mathcal{X} \times \mathcal{X}} M_{i, \delta, \bar{\delta}'} - \sum_{\bar{\delta} \in \mathcal{X} \times \mathcal{X}} M_{i, \bar{\delta}, \delta} = \sum_{\bar{\delta}': \delta \sqsupseteq_i \bar{\delta}'} q_{i, \delta, \bar{\delta}'} - \sum_{\bar{\delta}: \bar{\delta} \sqsupseteq_i \delta} q_{i, \bar{\delta}, \delta} \qquad \text{(by construction of } M)$$

$$= (\boldsymbol{q}^T Q)_{i\delta} \qquad \text{(by definition of } Q)$$

$$= -(\boldsymbol{p}^T X)_{i\delta} \qquad \text{(by condition 2)}$$

$$= -\sum_{j: (\boldsymbol{x}^j, \boldsymbol{z}^j) = \delta} p_j x_i^j + \sum_{j: (\boldsymbol{x}'^j, \boldsymbol{z}'^j) = \delta} p_j x_i'^j \qquad \text{(by definition of } X \text{ and } \kappa_i = 1)$$

$$= -\sum_{j: (\boldsymbol{x}^j, \boldsymbol{z}^j) = \delta} s_j x_i^j + \sum_{j: (\boldsymbol{x}'^j, \boldsymbol{z}'^j) = \delta} s_j x_i'^j \qquad \text{(by construction of } \boldsymbol{s})$$

$$= -c_{i, \delta} \qquad \text{(by definition of } c_{i, \delta})$$

So losses influence-dominate wins when $\kappa_i = 1$. The same argument will show that when $\kappa_i = -1$, wins influence-dominate losses. $\square$

References

Alesina, A., M. Carlana, E. L. Ferrara, and P. Pinotti (2018). Revealing Stereotypes: Evidence from Immigrants in Schools. *NBER Working Paper 25333*.

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.

Arrow, K. J. (1973). The Theory of Discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton University Press.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2022). Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment. *CESifo Working Paper 9731*.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

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.

Bodner, R. and D. Prelec (2003). Self-signaling and Diagnostic Utility in Everyday Decision Making. In I. Brocas and J. D. Carrillo (Eds.), *The Psychology of Economic Decisions Volume One: Rationality and Well-Being*. Oxford: Oxford University Press.

Bohnet, I., A. van Geen, and M. Bazerman (2016). When Performance Trumps Gender Bias: Joint vs. Separate Evaluation. *Management Science 62*(5), 1225–1234.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2021). Inaccurate Statistical Discrimination: An Identification Problem. *mimeo*.

Bohren, J. A., P. Hull, and A. Imas (2022). Systemic Discrimination: Theory and Measurement. *NBER Working Paper 29820*.

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

Border, K. C. (2013). Alternative Linear Inequalities, version 2020.10.15::09.50, accessed 2022-04-24. *https://kcborder.caltech.edu/Notes/Alternative.pdf*.

Bursztyn, L., G. Egorov, I. Haaland, A. Rao, and C. Roth (2022). Justifying Dissent. *NBER Working Paper 29730*.

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.

Caruso, E. M., D. A. Rahnev, and M. R. Banaji (2009). Using Conjoint Analysis to Detect Discrimination: Revealing Covert Preferences From Overt Choices. *Social Cognition 27*(1), 128–137.

Chambers, C. P. and F. Echenique (2016). *Revealed Preference Theory*. Econometric Society Monographs (56). Cambridge University Press.

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.

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. *mimeo*.

Cunningham, T. (2016). Hierarchical Aggregation of Information and Decision-Making. *mimeo*.

Dana, J., D. M. Cain, and R. M. Dawes (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes 100*(2), 193–201.

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.

de Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and Bounding Experimenter Demand. *American Economic Review 108*(11), 3266–3302.

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.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology 56*(1), 5–18.

Ellis, A. and Y. Masatlioglu (2021). Choice with Endogenous Categorization. *The Review of Economic Studies 89*(1), 240–278.

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.

Hodson, G., J. F. Dovidio, and S. L. Gaertner (2002). Processes in Racial Discrimination: Differential Weighting of Conflicting Information. *Personality and Social Psychology Bulletin 28*(4), 460–471.

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.

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.

Mummolo, J. and E. Peterson (2018). Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review 113*(2), 517–529.

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.

Phelps, E. (1972). The Statistical Theory of Racism and Sexism. *American Economic Review 62*(4), 659–61.

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.

Reuben, E., P. Sapienza, and L. Zingales (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences 111*(12), 4403–4408.

Ridout, S. (2021). Choosing for the Right Reasons. *mimeo*.

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics 17*(3), 523–534.

Simonsohn, U. (2016). DataColada[51] Greg vs. Jamal: Why Didn't Bertrand and Mullainathan (2004) Replicate? *https://datacolada.org/51, accessed 2022-05-01*.

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.

Uhlmann, E. and G. L. Cohen (2005). Constructed Criteria: Redefining Merit to Justify Discrimination. *Psychological Science 16*(6), 474–80.

Woodford, M. (2012). Inattentive Valuation and Reference-Dependent Choice. *mimeo*.

A   Web appendix to "Implicit Preferences"

For Online Publication Only

A.1   Derivations for Section 3

We use two tricks to simplify the derivation of the corollaries. First we show that we can represent each of the cases with just three attributes without loss of generality. Second we represent each dataset with a single row of a matrix, using a compressed version of the matrix X that we derived in the proof of Theorem 1. With this row representation most derivations become simple: we can see which realizations of $\boldsymbol{\kappa}$ are infeasible by observing whether some combination of rows in Q (i.e. combination of $\sqsupseteq_i$ relations) rules them out.

Recall that we assume Equivalence throughout, which implies that comparisons with the same difference $|\boldsymbol{x} - \boldsymbol{z}|$ have the same influence $\theta_i, \forall i$.

Reduction to three attributes. All of our examples can be analyzed by partitioning the set of attributes into three disjoint and collectively exhaustive "groups," A, B, C, where all attributes within a group are perfectly correlated, so we can represent them using three grouped attributes, $\boldsymbol{x} = (x_A, x_B, x_C)$.[38] Since attributes are perfectly correlated within groups, they will have identical differences in a given comparison (e.g. we have $|x_i - z_i| = |x_j - z_j|, \forall i, j \in A$), and all influence-dominance relationships will be shared. So, for example $\big((\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_i (\boldsymbol{x}', \boldsymbol{z}')\big) \Leftrightarrow \big((\boldsymbol{x}, \boldsymbol{z}) \sqsupseteq_j (\boldsymbol{x}', \boldsymbol{z}')\big), \forall i, j \in A$. Therefore we can conduct all our analysis using x_A, x_B, x_C, where $x_A \kappa_A \theta_A := \sum_{i \in A} x_i \kappa_i \theta_i$. Implications that we derive on a grouped attribute will imply a disjunction over all attributes within the group (essentially, because we do not know which attribute(s) within a group are responsible for the observed behavior). That is: $(x_A \kappa_A = 1) \implies \big(\bigvee_{i \in A} x_i \kappa_i = 1\big).$

Applying Theorem 1 compactly. The proof of Theorem 1 shows how to represent a dataset and influence-dominance relationship in terms of P, X, and Q matrices, and use them to ask whether a given $\boldsymbol{\kappa}$ can rationalize the data. Condition 2 of the theorem tells us the answer is no if and only if there exist vectors $\boldsymbol{p}, \boldsymbol{q}$ such that $\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} P & X \\ 0 & Q \end{bmatrix} = \boldsymbol{0}$. Condition 3 tells us that $\boldsymbol{p}$ is a cyclical selection and $\boldsymbol{q}$ is a matching.

In order to parsimoniously identify *every* $\boldsymbol{\kappa}$ that can be ruled out in this way, we will write out the terms of the expression for an arbitrary $\boldsymbol{\kappa}$, and ask for which κ_i values least one term must be nonzero. However, in general the matrices can be very large. We use a number of tricks to substantially compress them:

1. We can ignore P, since in any solution, $\boldsymbol{p}$ is a cyclical selection and P's rows always

---

[38]So, $A \cup B \cup C = 1, \ldots, n$; $A \neq \emptyset, B \neq \emptyset, C \neq \emptyset$; and $A \cap B = A \cap C = B \cap C = \emptyset$. For example, if $A = \{1, 2, 3\}$ we might have $x_A = -1 \Leftrightarrow (x_1, x_2, x_3) = (-1, 1, -1)$ and $x_A = 1 \Leftrightarrow (x_1, x_2, x_3) = (1, -1, 1)$.

sum to zero in a cyclical selection. Thus we focus on X and Q.

2. When the dataset consists of a single intransitive choice cycle, we can reduce its X matrix to a single row by summing the individual rows. This is because $\boldsymbol{p}$ is a cyclical selection and in a cyclical selection every bundle must appear equally often on the LHS and RHS, meaning all p_j terms must be equal.

3. As defined, X and Q have many columns, indexed by i, and $(\boldsymbol{x}, \boldsymbol{z})$. However, for each i we can without loss of generality add together columns with identical differences $(|\boldsymbol{x} - \boldsymbol{z}| = |\boldsymbol{x}' - \boldsymbol{z}'|)$. Equivalence tells us they must have the same θ_i, and all that matters for establishing whether wins have higher influence than losses is the net wins or losses for each *distinct* realization of θ_i. (The intuition is the same as for why it is sufficient to match *scores* rather than each individual win or loss). We therefore write X and Q with one column per realization of $|\boldsymbol{x} - \boldsymbol{z}|$. We construct X by counting wins and losses for each i and $|\boldsymbol{x} - \boldsymbol{z}|$, and we construct Q from the restrictions implied by Assumptions 2 and 3.

4. Finally, many of the potential comparisons in $\mathcal{X} \times \mathcal{X}$ are never observed, so appear in X as columns of zeros. We can ignore those columns. Similarly, Q will have many rows that do not restrict any nonzero column in X. We eliminate those as well.

We name the compressed X and Q matrices X^* and Q^* and show them in Figure 4. Most corollaries are easily verified by visual inspection of the Figure; we recommend the reader use the matrix as a rubric to understand the results. One needs only to confirm that the representation in X^* is accurate, and then observe which combination of $\boldsymbol{\kappa}$ would be ruled out by some combination of rows in Q^*. We include full proofs below for completeness.

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 non-shared attributes in $|\boldsymbol{x}^1 - \boldsymbol{x}^2|$, B is the set of non-shared attributes in $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$, $A \cup B$ is the set of non-shared attributes in $|\boldsymbol{x}^1 - \boldsymbol{x}^3|$ (the "diagonal"), and C is the set that are always shared. Because $|\boldsymbol{x}^1 - \boldsymbol{x}^2|$ and $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$ differ on distinct attributes, A, B, C are disjoint and collectively exhaustive. We have:

$$|\boldsymbol{x}^1 - \boldsymbol{x}^2| = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, |\boldsymbol{x}^2 - \boldsymbol{x}^3| = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \text{ and } |\boldsymbol{x}^1 - \boldsymbol{x}^3| = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}.$$

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 wins and losses for each i and $|\boldsymbol{x} - \boldsymbol{z}|$. 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

50

Top panel — $X^* = $

	$A,\left[\begin{smallmatrix}2\\0\end{smallmatrix}\right]$	$A,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$A,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	$B,\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$	$B,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$B,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	$C,\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$	$C,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$C,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$
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
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
Figure 8	$-4\kappa_A x_A^4$	0	$4\kappa_A x_A^4$	0	0	0	0	0	0
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	$\kappa_C x_C \Upsilon$
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	$\kappa_C \bar{x}_C \bar{\Upsilon}$

$Q^* = $

	$A,\left[\begin{smallmatrix}2\\0\end{smallmatrix}\right]$	$A,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$A,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	$B,\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$	$B,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$B,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	$C,\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$	$C,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$C,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$
$\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]\sqsupseteq_A\left[\begin{smallmatrix}2\\0\end{smallmatrix}\right]$	-1	0	1	0	0	0	0	0	0
$\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]\sqsupseteq_B\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	0	0	0	-1	0	1	0	0	0
$\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]\sqsupseteq_C\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	0	0	0	0	0	0	1	0	-1
Dominance	0	0	0	Θ	0	$-\Theta$	0	0	0

Bottom panel — $X^* = $

	$1,\left[\begin{smallmatrix}2\\0\end{smallmatrix}\right]$	$1,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$1,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	$2,\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$	$2,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$2,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$	$3,\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$	$3,\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$	$3,\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$
Equilateral triangle	0	$-2\kappa_1$	$2\kappa_1$	0	$-2\kappa_2$	$2\kappa_2$	0	$-2\kappa_3$	$2\kappa_3$
Non-convex scissor	$-\kappa_1$	κ_1	0	κ_2	$-\kappa_2$	0	κ_3	$-\kappa_3$	0
Falsification	0	0	0	0	0	0	0	0	0

$Q^* = $ empty

Figure 4: Matrix representation of corollaries and examples from Section 3

Top panel corresponds to the main Corollaries. (1) Columns are labeled by attribute group ($i \in \{A,B,C\}$), and $|x-z| \in \{0,2\}^3$. (2) Rows correspond to κ_i multiplied by net wins and losses for that example. (3) Q^* includes only rows that restrict at least one row of X^*. (4) For scissors, $\Upsilon \in \{-1,1\}$ equals the sign of the evaluation change: $\Upsilon = sgn(y^2 - y^1)$. (5) $\Theta \in \{-1,0,1\}$ captures the sign of the Dominance of attribute k assumption (Assumption 3). $\Theta = 0$ if the assumption does not apply, $\Theta = 1$ if influence is higher for shared attributes (k is shared), $\Theta = -1$ if influence is higher for non-shared (k is non-shared). **Bottom panel** shows the additional cases discussed under "Other examples." Attributes 1, 2, 3 correspond to horizontal, vertical, depth. Q^* is empty as $\sqsupseteq$ does not restrict any row of X^* for these examples.

otherwise. From the conditions defining the right triangle, we know that $x_A^1 = -x_A^2 = -x_A^3$ while $x_B^1 = x_B^2 = -x_B^3$ and $x_C^1 = x_C^2 = x_C^3$. We work through each inequality in turn.

Inequality 1 gives us two wins for A, $\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$ if $x_A^1 = 1$ and two losses if $x_A^1 = -1$. Thus the entry in column A, $\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$ equals $2\kappa_A x_A^1$, which in turn equals $-2\kappa_A x_A^3$ by definition of x^1 and x^3. All other attributes are shared so have zero net wins or losses (each win on the LHS is canceled by a loss on the RHS and vice versa).

Inequality 2 gives us two wins for B, $\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$ if $x_B^2 = 1$ and two losses if $x_B^2 = -1$. Thus the entry in column B, $\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$ equals $2\kappa_B x_B^2$, which in turn equals $-2\kappa_B x_B^3$ by definition of x^2 and x^3. All other attributes do not vary and so have zero net wins or losses.

Inequality 3 gives us two wins for A, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ if $x_A^3 = 1$ and two losses if $x_A^3 = -1$. Thus the entry in column A, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ equals $2\kappa_A x_A^3$. Inequality 3 gives us two wins for B, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ if $x_B^3 = 1$ and two losses if $x_B^3 = -1$. Thus the entry in column B, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ equals $2\kappa_B x_B^3$. All other attributes do not vary and so have zero net wins or losses.

Collapsing these entries to a single row we obtain "Right triangle 1" in Figure 4.

Because our dataset consists of a single cycle we can set $\boldsymbol{p} = 1$ without loss of generality, obtaining (after ignoring columns that equal zero):

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

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: $(\kappa_A x_A^3 \leq 0) \wedge (\kappa_B x_B^3 \leq 0)$. Hence, the data can be rationalized if and only if:

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

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 non-shared attributes in $|\boldsymbol{x}^1 - \boldsymbol{x}^2|$ and $|\boldsymbol{x}^3 - \boldsymbol{x}^4|$, B is the set of *additional* attributes that are non-shared in $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$ and $|\boldsymbol{x}^1 - \boldsymbol{x}^4|$ but were shared in $|\boldsymbol{x}^1 - \boldsymbol{x}^2|$ and $|\boldsymbol{x}^3 - \boldsymbol{x}^4|$, $A \cup B$ the set of all attributes that are non-shared in $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$ and $|\boldsymbol{x}^1 - \boldsymbol{x}^4|$, and C the set that are shared in all comparisons. By construction, A, B, C are disjoint and collectively exhaustive. We have:

$$|\boldsymbol{x}^1 - \boldsymbol{x}^2| = |\boldsymbol{x}^3 - \boldsymbol{x}^4| = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \text{ and } |\boldsymbol{x}^2 - \boldsymbol{x}^3| = |\boldsymbol{x}^1 - \boldsymbol{x}^4| = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}.$$

As with the right triangle, we populate the matrix X^* by calculating wins and losses for each

$i, |\boldsymbol{x} - \boldsymbol{z}|$ combination. In this case all comparisons are concentrated on just two $|\boldsymbol{x} - \boldsymbol{z}|$'s. Following the same proof strategy as for the right triangle, we set $\boldsymbol{p} = 1$ without loss of generality. We obtain (after eliminating zeros):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X^* \\ Q^* \end{bmatrix} = \begin{bmatrix} -4\kappa_A x_A^4 - q_1 & 4\kappa_A x_A^4 + q_1 \end{bmatrix},$$

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:

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

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 $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$ and not shared in $|\bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2|$, and C is the set of attributes that are always shared within any comparison.[39] By construction, A, B, and C are disjoint and collectively exhaustive.

We populate the second triangle's row in X^* by calculating the wins and losses for each $i, |\bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}|$ combination. As for right triangle 1 we exploit 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 (ignoring zeros):

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

where q_1 is the coefficient on the first row of Q^* and q_2 is the coefficient on the second. The

---

[39]Note that while C attributes do not vary within any comparisons, they might differ *between* the two triangles, that is it could be that $\boldsymbol{x}_C \neq \bar{\boldsymbol{x}}_C$.

second steps use the fact that $x_A^3 = -\bar{x}_A^3$, and $x_B^3 = \bar{x}_B^3$.[40] 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 cyclical selection contains an equal number of each cycle), the disjunction collapses to $(\kappa_B(p_1 + p_2)x_B^3 > 0)$, 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 $\boldsymbol{p}$s. Finally, using the definition of set B we obtain the result, that a pair of parallel right triangles implies:

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

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, $A \cup B$ the full set that vary in the second comparison, and C the set that do not vary within either comparison. 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 $y^2 > y^1$ we have $u(\boldsymbol{x}, \boldsymbol{z}^2) > u(\boldsymbol{x}, \boldsymbol{z}^1)$. The left-hand side corresponds to $|\boldsymbol{x} - \boldsymbol{z}^2| = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$, giving us a win in column i, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ if $x_i = 1$ and a loss otherwise. The right-hand side corresponds to $|\boldsymbol{x} - \boldsymbol{z}^1| = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$, giving us a loss in column i, $\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$ if $x_i = 1$ and a win otherwise. If $y^2 < y^1$ then the left- and right-hand sides of the inequalities are switched. Thus, defining $\Upsilon = sgn(y^2 - y^1)$, we enter $\kappa_i x_i \Upsilon$ in the columns associated with difference $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$, and $-\kappa_i x_i \Upsilon$ in the columns associated with difference $\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$. Thus, we obtain

---

[40]The definition of the parallel right triangle, condition (1) $(\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2)$ allows us to pin down the values of the non-shared attributes in these comparisons (set B): $(\boldsymbol{x}_B^2 = \bar{\boldsymbol{x}}_B^1)$ and $(\boldsymbol{x}_B^3 = \bar{\boldsymbol{x}}_B^2)$ (to see this note that if $\boldsymbol{x}_B^2 - \boldsymbol{x}_B^3 = 2$, it must be that $\boldsymbol{x}_B^2 = 1$ and $\boldsymbol{x}_B^3 = -1$). Similarly, condition (2) $(\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3))$ allows us to pin down the values of the non-shared attributes in these comparisons (set A): $(\boldsymbol{x}_A^1 = -\bar{\boldsymbol{x}}_A^2)$ and $(\boldsymbol{x}_A^2 = -\bar{\boldsymbol{x}}_A^3)$. 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 $x_B^3 = \bar{x}_B^3$.

(setting $p_1 = 1$ and ignoring zeros):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X^* \\ Q^* \end{bmatrix} = \begin{bmatrix} -W_A & W_A & -W_B & W_B & -W_C & W_C \end{bmatrix}$$

$$W_A = \kappa_A x_A \Upsilon + q_1$$

$$W_B = \kappa_B x_B \Upsilon - \Theta q_4$$

$$W_C = \kappa_C x_C \Upsilon - q_3.$$

q_1 and q_3 are the coefficients on the first and third rows of Q^*, which correspond to the Dilution assumption (Assumption 2), while q_4 is the coefficient on Q^*'s fourth row which captures the Dominance of attribute k assumption (Assumption 3). Θ encodes Assumption 3: $\Theta = 0$ if the assumption does not apply, $\Theta = 1$ if influence is higher for shared attributes (k is shared), $\Theta = -1$ if influence is higher for non-shared attributes (k is non-shared).

When $\Theta = 0$, the data can be rationalized if and only if there is no $\boldsymbol{q} \geq 0$ such that $W_A = W_B = W_C = 0$, i.e. if and only if:

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

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

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

Expanding these expressions using the definitions of A, B, C, Υ and Θ gives the results.

Parallel convex scissors without and with Dominance of attribute k. The conditions (1) and (2) imply $|\boldsymbol{x} - \boldsymbol{y}^1| = |\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}^1|$ and $|\boldsymbol{x} - \boldsymbol{y}^2| = |\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}^2|$. Let:

$$A = \{i : x_i \neq z_i^1\} = \{i : \bar{x}_i \neq \bar{z}_i^1\}$$

$$B = \{i : z_i^1 \neq z_i^2\} = \{i : \bar{z}_i^1 \neq \bar{z}_i^2\}$$

$$C = \{i : x_i = z_i^2\} = \{i : \bar{x}_i = \bar{z}_i^2\}.$$

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 within any comparison. By construction, A, B, C are disjoint and collectively exhaustive. Since the values of $\boldsymbol{x}, \bar{\boldsymbol{x}}, sgn(y^2 - y^1)$ and $sgn(\bar{y}^2 - \bar{y}^1)$ are unrestricted, there are many possible combinations.

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 (ignoring zero elements):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \begin{bmatrix} X^* \\ Q^* \end{bmatrix} = \begin{bmatrix} -W_A & W_A & -W_B & W_B & -W_C & W_C \end{bmatrix}$$

$$W_A = \kappa_A(p_1 x_A \Upsilon + p_2 \bar{x}_A \bar{\Upsilon}) + q_1$$

$$W_2 = \kappa_B(p_1 x_B \Upsilon + p_2 \bar{x}_B \bar{\Upsilon}) - \Theta q_4$$

$$W_3 = \kappa_C(p_1 x_C \Upsilon + p_2 \bar{x}_C \bar{\Upsilon}) - q_3,$$

where q_1 and q_3 are the coefficients on the first and third rows of Q^*, which capture the Dilution assumption (Assumption 2), while q_4 is the coefficient on Q^*'s fourth row which captures Dominance of attribute k (Assumption 3) as before. Θ encodes Assumption 3. $\Upsilon = sgn(y^2 - y^1)$ and $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^1)$ capture the direction in which each evaluation changes when the comparator changes.

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 cyclical selection consisting of exactly one copy of each scissor ($p_1 = p_2 = 1$).

When $\Theta = 0$ the data can be rationalized if and only if:

$$\left(\kappa_A(x_A \Upsilon + \bar{x}_A \bar{\Upsilon}) = 2 \right) \vee \left(\kappa_B(x_B \Upsilon + \bar{x}_B \bar{\Upsilon}) \neq 0 \right) \vee \left(\kappa_C(x_C \Upsilon + \bar{x}_C \bar{\Upsilon}) = -2 \right).$$

When $\Theta \neq 0$ the data can be rationalized if and only if:

$$\left(\kappa_A(x_A \Upsilon + \bar{x}_A \bar{\Upsilon}) = 2 \right) \vee \left(\kappa_B(x_B \Upsilon + \bar{x}_B \bar{\Upsilon}) = -2\Theta \right) \vee \left(\kappa_C(x_C \Upsilon + \bar{x}_C \bar{\Upsilon}) = -2 \right).$$

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 also that some parallel scissors eliminate the term involving attribute group B (the attributes about which a single scissor is indeterminate), enabling us to draw precise conclusions without invoking Assumption 3.

A.1.1 Corollary 3 is sufficient but not necessary for falsification

Corollary 3 says that to falsify the model it is sufficient but not necessary that the dataset D contains a cyclical selection where losses influence-dominate wins and wins influence-dominate losses. We show this condition is not necessary by providing a counterexample.

Consider a decision maker who satisfies Assumptions 1 and 2, choosing between bundles with $n = 4$ attributes. On the left side of the diagram below we draw two figure-8 cycles, in three dimensions, holding the fourth fixed. The first figure-8 has $x_4 = -1$ and the second figure-8 has $x_4 = 1$. The first cycle rules out all $\boldsymbol{\kappa}$s with $\kappa_2 \neq 1$, while the second rules out all $\boldsymbol{\kappa}$s with $\kappa_2 \neq -1$, so there exists no $u^I(\boldsymbol{x}, \boldsymbol{z})$ that can rationalize the dataset. On the right of the diagram we show the simplified matrix representation of the dataset (see "Applying Theorem 1 compactly" above, note that only attribute 2 has nonzero columns in X^*). Observe that no single cyclical selection (no weighted combination of rows of X^*) can be matched to rows of Q^* to obtain a row of zeroes.

$$2, \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \quad 2, \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} \quad 2, \begin{bmatrix} 0 \\ 2 \\ 2 \\ 0 \end{bmatrix}$$

$$X^* = \begin{matrix} \text{1st Figure-8} \\ \text{2nd Figure-8} \end{matrix} \begin{bmatrix} -4\kappa_2 & 4\kappa_2 & 0 \\ 4\kappa_2 & 0 & -4\kappa_2 \end{bmatrix}$$

$$Q^* = \begin{matrix} \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix} \sqsupseteq_2 \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 2 \\ 2 \\ 0 \end{bmatrix} \sqsupseteq_2 \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \end{matrix} \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

A.2 Proofs for Section 4 (Foundations)

In proving some of these results we make use of an additional lemma that we call "Sums and Differences," which we state and prove first. Recall also the definition of the set of shared attributes: $S^{(\boldsymbol{x}, \boldsymbol{z})} = \{i : |x_i - z_i| = 0\}$.

Lemma 2 (Sums and Differences). *Suppose we observe two linear combinations of n independent Normal variables ("weights"), with $+1$ or -1 coefficients ("attributes"):*

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

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

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

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

The posterior for the weight on a shared attribute depends only on the sum $\bar{y}^x + \bar{y}^z$, and the posterior for the weight on a non-shared attribute depends only on the difference $\bar{y}^x - \bar{y}^z$.

Proof of Lemma 2. First we assume there exists at least one shared and one non-shared attribute (i.e., $\boldsymbol{x} \neq \boldsymbol{z}$ and $\boldsymbol{x} \neq -\boldsymbol{z}$). Given two multivariate Normals, $\boldsymbol{a}$ and $\boldsymbol{b}$, with covariance $Var\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \Sigma_a & \Sigma_{a,b} \\ \Sigma_{a,b}^T & \Sigma_b \end{bmatrix}$ 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:

$$E[\boldsymbol{w}|\boldsymbol{y}] = \Sigma_{w,y}\Sigma_y^{-1}\boldsymbol{y} \tag{4}$$

with components as follows:

$$\Sigma_y = X\Sigma_w X^T = \begin{bmatrix} \sum_i x_i^2 \sigma_i^2 & \sum_i x_i z_i \sigma_i^2 \\ \sum_i x_i z_i \sigma_i^2 & \sum_i z_i^2 \sigma_i^2 \end{bmatrix} = \begin{bmatrix} \sum_{i \in S} \sigma_i^2 + \sum_{i \notin S} \sigma_i^2 & \sum_{i \in S} \sigma_i^2 - \sum_{i \notin S} \sigma_i^2 \\ \sum_{i \in S} \sigma_i^2 - \sum_{i \notin S} \sigma_i^2 & \sum_{i \in S} \sigma_i^2 + \sum_{i \notin S} \sigma_i^2 \end{bmatrix}$$

$$\Sigma_y^{-1}\boldsymbol{y} = \frac{1}{4\sum_{i \in S}\sigma_i^2 \sum_{i \notin S}\sigma_i^2} \begin{bmatrix} \left(\sum_{i \in S}\sigma_i^2 + \sum_{i \notin S}\sigma_i^2\right)\bar{y}^x + \left(-\sum_{i \in S}\sigma_i^2 + \sum_{i \notin S}\sigma_i^2\right)\bar{y}^z \\ \left(-\sum_{i \in S}\sigma_i^2 + \sum_{i \notin S}\sigma_i^2\right)\bar{y}^x + \left(\sum_{i \in S} +\sigma_i^2 \sum_{i \notin S}\sigma_i^2\right)\bar{y}^z \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} \dfrac{\bar{y}^x + \bar{y}^z}{\sum_{i \in S}\sigma_i^2} + \dfrac{\bar{y}^x - \bar{y}^z}{\sum_{i \notin S}\sigma_i^2} \\ \dfrac{\bar{y}^x + \bar{y}^z}{\sum_{i \in S}\sigma_i^2} - \dfrac{\bar{y}^x - \bar{y}^z}{\sum_{i \notin S}\sigma_i^2} \end{bmatrix}$$

$$\Sigma_{w,y} = \Sigma_w X^T = \begin{bmatrix} x_1 \sigma_1^2 & z_1 \sigma_1^2 \\ \vdots & \vdots \\ x_n \sigma_n^2 & z_n \sigma_n^2 \end{bmatrix}$$

Thus, given (4), we obtain:

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

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

The same formula applies to the two special cases we initially ruled out, $\boldsymbol{x} = \boldsymbol{z}$ and $\boldsymbol{x} = -\boldsymbol{z}$. We cannot use equation (4) because X does not have full rank so Σ_y is not

invertible. If all attributes are shared ($\boldsymbol{x} = \boldsymbol{z}$) we have a Normal updating problem with a single observable, $\bar{y}^x = \bar{y}^z$, and each w_i is updated in proportion to its share of the total variance. So, $E[w_i|\boldsymbol{y}] = x_i \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2} \bar{y}^x = x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\bar{y}^x + \bar{y}^z}{2}$. If all attributes are non-shared ($\boldsymbol{x} = -\boldsymbol{z}$) then $\bar{y}^x = -\bar{y}^z$ and we have $E[w_i|\boldsymbol{y}] = x_i \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2} \bar{y}^x = x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\bar{y}^x - \bar{y}^z}{2}$. Thus both correspond to the statement of the Lemma. $\square$

Proof Strategy for Propositions 1–5. To prove Propositions 1–4 we need to show utility function defined in each foundation can be expressed as an Implicit Preferences utility function (consistent with equation (1)), where $\theta_i(\boldsymbol{x}, \boldsymbol{z})$ depends only on $|\boldsymbol{x} - \boldsymbol{z}|$ (Equivalence, Assumption 1) and weakly increases as more attributes share status with i (Dilution, Assumption 2). To verify Equivalence we will show that in each foundation, $\theta_i(\boldsymbol{x}, \boldsymbol{z})$ can be written as:

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

To verify Dilution we show that $\theta_i(\boldsymbol{x}, \boldsymbol{z})$ is weakly increasing as the set of attributes that share status with i grows. We can study the properties of θ_i^S and θ_i^N separately, since i does not change status in a given dilution. We therefore 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.

To prove Proposition 5, for each foundation we write out $\theta_i(\boldsymbol{x}, \boldsymbol{z}) - \theta_i(\boldsymbol{x}', \boldsymbol{z}')$ and show that the conditions of the Proposition imply it is weakly positive.

Proof of Proposition 1. First, note that $u^{CP}(\boldsymbol{x}, \boldsymbol{z})$ can be rearranged to satisfy equation (1) (using the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$):

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

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

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 (rules are only applied to non-shared attributes). We need to show that θ_i^N is weakly increasing as the set of non-shared attributes grows. Let $(\boldsymbol{x}', \boldsymbol{z}')$ be a dilution of $(\boldsymbol{x}, \boldsymbol{z})$ with respect to attribute i. 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 $\theta_i^N(|\boldsymbol{x}' - \boldsymbol{z}'|) = |\lambda_i| \geq \theta_i^N(|\boldsymbol{x} - \boldsymbol{z}|)$. $\qquad\square$

Proof of Proposition 2. First we derive an explicit solution for the observer's posterior.

Lemma 3. *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.

Proof of Lemma 3. 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 $(x_i - z_i)^2 = 4 \times \mathbf{1}\{i \notin S\}$. $\qquad\square$

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.

Using Lemma 3 and the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$, we can rearrange u^{SC} to satisfy (1):

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

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

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. $\qquad\square$

Next, we show that reporting u^{SE} is an optimal strategy in the signaling-evaluation game:

Proof of Lemma 1. Define the *residual* evaluations $\bar{y}^x, \bar{y}^z$, after subtracting components which are common knowledge. We have:

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

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 the optimal reports:

$$y^x(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i w_i + \sum_{i=1}^{n} \lambda_i \frac{\partial \hat{w}_i(y^x, y^z)}{\partial y^x}$$

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

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

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. Taking first order conditions of U^2, and using Lemma 2, we obtain the desired result:

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

$\square$

Proof of Proposition 3. Using the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$ we can rearrange $u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ in a form that satisfies (1):

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

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

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

Proof of Proposition 4. First, we show that utility takes a simple form:

Lemma 4. *An implicit associations utility function can be written as:*

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

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

Proof of Lemma 4. Define $\hat{f}(\boldsymbol{x}) = E[f(\boldsymbol{x})|\boldsymbol{\lambda}]$. Given agent 1's prior on $\boldsymbol{\pi}$, we have:

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

Next, we define the residual value $\bar{\hat{f}}(\boldsymbol{x})$ by subtracting the common-knowledge $g(\boldsymbol{x})$. We obtain $\bar{\hat{f}}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \lambda_i$. The second agent's posteriors for each λ_i can then

be derived using Lemma 2:

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

$$= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \frac{\sum_{i=1}^{n}(x_i+z_i)\lambda_i}{2} & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \frac{\sum_{i=1}^{n}(x_i-z_i)\lambda_i}{2} & , i \notin S \end{cases}$$

$$= \begin{cases} x_i \frac{\sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} x_j \lambda_j & , i \in S \\ x_i \frac{\sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \sum_{j \notin S} x_j \lambda_j & , i \notin S \end{cases}$$

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

$$E[f(\boldsymbol{x})|\boldsymbol{\pi}, \hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})] = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \pi_i E[\lambda_i|\hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})]$$

$$= g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i^2 \pi_i \sigma_i^2 \left(\frac{\mathbf{1}\{i \in S\}}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} x_j \lambda_j + \frac{\mathbf{1}\{i \notin S\}}{\sum_{j \notin S} \sigma_j^2} \sum_{j \notin S} x_j \lambda_j \right)$$

$$= g(\boldsymbol{x}) + \frac{\sum_{i \in S} \pi_i \sigma_i^2}{\sum_{j \in S} \sigma_j^2} \sum_{j \in S} x_j \lambda_j + \frac{\sum_{i \notin S} \pi_i \sigma_i^2}{\sum_{j \notin S} \sigma_j^2} \sum_{j \notin S} x_j \lambda_j$$

$$= g(\boldsymbol{x}) + \sum_{i \in S} x_i \lambda_i \frac{\sum_{j \in S} \pi_j \sigma_j^2}{\sum_{j \in S} \sigma_j^2} + \sum_{i \notin S} x_i \lambda_i \frac{\sum_{j \notin S} \pi_j \sigma_j^2}{\sum_{j \notin S} \sigma_j^2} = u^{IA}(\boldsymbol{x}, \boldsymbol{z}),$$

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 fourth step uses a switch of index labels. $\square$

Now we are ready to prove the Proposition. It states that at most one attribute has either $\lambda_i \neq 0$ or $\pi_i \neq 1$. Assign index t to this attribute. Our first goal is to show that the functional form derived in Lemma 4 can be written in a form satisfying (1). Observe that

$\bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|)$ can be written as;

$$
\bar{\pi}_i(|\boldsymbol{x} - \boldsymbol{z}|) =
\begin{cases}
1 - \frac{\sum_{j \in S}(1 - \pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} & , i \in S \\
1 - \frac{\sum_{j \notin S}(1 - \pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2} & , i \notin S
\end{cases}
$$

$$
=
\begin{cases}
1 - \frac{(1 - \pi_i)\sigma_i^2}{\sum_{j \in S}\sigma_j^2} - \frac{\sum_{(j \in S) \wedge (j \neq i)}(1 - \pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} & , i \in S \\
1 - \frac{(1 - \pi_i)\sigma_i^2}{\sum_{j \notin S}\sigma_j^2} - \frac{\sum_{(j \notin S) \wedge (j \neq i)}(1 - \pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2} & , i \notin S
\end{cases}
$$

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

Substituting into the functional form derived in Lemma 4, and using the fact that $\lambda_i(1 - \pi_i) = sgn(\lambda_i(1 - \pi_i))|\lambda_i(1 - \pi_i)|$, we obtain:

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

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

where:

$$
B = \sum_{i=1}^{n} x_i \lambda_i \left[\mathbf{1}\{i \in S\}\frac{\sum_{(j \in S) \wedge (j \neq i)}(1 - \pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} + \mathbf{1}\{i \notin S\}\frac{\sum_{(j \notin S) \wedge (j \neq i)}(1 - \pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2}\right]
$$

$$
= x_t \lambda_t \left[\mathbf{1}\{t \in S\}\frac{\sum_{(j \in S) \wedge (j \neq t)}(1 - \pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} + \mathbf{1}\{t \notin S\}\frac{\sum_{(j \notin S) \wedge (j \neq t)}(1 - \pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2}\right]
$$

$$
+ \sum_{i \neq t} x_i \lambda_i \left[\mathbf{1}\{i \in S\}\frac{\sum_{(j \in S) \wedge (j \neq i)}(1 - \pi_j)\sigma_j^2}{\sum_{j \in S}\sigma_j^2} + \mathbf{1}\{i \notin S\}\frac{\sum_{(j \notin S) \wedge (j \neq i)}(1 - \pi_j)\sigma_j^2}{\sum_{j \notin S}\sigma_j^2}\right]
$$

$$
= 0,
$$

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). Thus we can see that u^{IA} is an Implicit Preferences utility function (satisfies equation (1)).

Finally, observe that θ_i depends only on $(\boldsymbol{x}, \boldsymbol{z})$ through $|\boldsymbol{x} - \boldsymbol{z}|$, and that θ^S and θ^N are weakly increasing as the sets of shared and non-shared attributes grow, respectively. $\qquad \square$

Proof of Proposition 5. Given the conditions in the Proposition, we need to show, for two comparisons $(\boldsymbol{x}, \boldsymbol{z})$ and $(\boldsymbol{x}', \boldsymbol{z}')$, where i and k share status in $(\boldsymbol{x}, \boldsymbol{z})$, and do not share status in $(\boldsymbol{x}', \boldsymbol{z}')$, that $\theta_i(\boldsymbol{x}, \boldsymbol{z}) \geq \theta_i(\boldsymbol{x}', \boldsymbol{z}')$.

Ceteris Paribus. $\theta_i(\boldsymbol{x}, \boldsymbol{z})$ can be written as:

$$\theta_i(\boldsymbol{x}, \boldsymbol{z}) = |\lambda_i|\big(1 - \mathbf{1}\{i \notin S^{(\boldsymbol{x}, \boldsymbol{z})}\}\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S^{(\boldsymbol{x}, \boldsymbol{z})})\}\big)$$

Observe that (1) $(i \in S^{(\boldsymbol{x}, \boldsymbol{z})}) \Rightarrow (\mathbf{1}\{i \notin S^{(\boldsymbol{x}, \boldsymbol{z})}\} = 0)$. (2) Since i and k share status in $(\boldsymbol{x}, \boldsymbol{z})$, we have that $(i \notin S^{(\boldsymbol{x}, \boldsymbol{z})}) \Rightarrow (k \notin S^{(\boldsymbol{x}, \boldsymbol{z})})$. Since by assumption $\lambda_k = 0$, $(k \notin S^{(\boldsymbol{x}, \boldsymbol{z})}) \Rightarrow (\mathbf{1}\{\forall j, (\lambda_j = 0) \Rightarrow (j \in S^{(\boldsymbol{x}, \boldsymbol{z})})\} = 0)$ (i.e., if k is non-shared, the rules are turned off, because k is not itself governed by a rule). Putting these together, we obtain that $\theta_i(\boldsymbol{x}, \boldsymbol{z}) = |\lambda_i|$. Therefore, we can write:

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

Signaling-evaluation and Implicit Associations. Both foundations have very similar influence functions. We can write:

$$\theta_i(\boldsymbol{x}, \boldsymbol{z}) - \theta_i(\boldsymbol{x}', \boldsymbol{z}') = A \times \sigma_i^2 \left(\frac{Z(\boldsymbol{x}, \boldsymbol{z}) - Z(\boldsymbol{x}', \boldsymbol{z}')}{Z(\boldsymbol{x}, \boldsymbol{z})Z(\boldsymbol{x}', \boldsymbol{z}')}\right)$$

where

$$Z(\boldsymbol{x}, \boldsymbol{z}) = \mathbf{1}\{i \in S^{(\boldsymbol{x}, \boldsymbol{z})}\} \sum_{j \in S^{(\boldsymbol{x}, \boldsymbol{z})}} \sigma_j^2 + \mathbf{1}\{i \notin S^{(\boldsymbol{x}, \boldsymbol{z})}\} \sum_{j \notin S^{(\boldsymbol{x}, \boldsymbol{z})}} \sigma_j^2$$

and where $A = |\lambda_i| \geq 0$ in the Signaling-evaluation model and $A = |\lambda_i(1 - \pi_i)| \geq 0$ in the Implicit Associations model. Observe that $Z(\boldsymbol{x}, \boldsymbol{z})Z(\boldsymbol{x}', \boldsymbol{z}') > 0$. Finally, observe that since k shares status with i in $(\boldsymbol{x}, \boldsymbol{z})$ and not in $(\boldsymbol{x}', \boldsymbol{z}')$, and by assumption $\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$, we must have $Z(\boldsymbol{x}, \boldsymbol{z}) - Z(\boldsymbol{x}', \boldsymbol{z}') \geq 0$. Hence $\theta_i(\boldsymbol{x}, \boldsymbol{z}) \geq \theta_i(\boldsymbol{x}', \boldsymbol{z}')$. $\square$

A.3 Appendix material for analysis of Exley (2016)

Data access: We accessed Exley's replication data through the journal webpage, at `https://doi.org/10.1093/restud/rdv051`.

A.3.1 Data structure

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 (or to another participant – we refer to both as "charity") that is chosen over $\$10$ for self.

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 $t = 0, \ldots, 20$. The sure payments are $Y_t^S = (0, 0.50, \ldots, 10)$ for self lotteries and $Y_t^C = (0, {}^X\!/_{20}, \ldots, X)$ 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.

A.3.2 Assumptions

Exley's null hypothesis, *standard risk preferences*, assumes two properties of utility. We will make use of the same assumptions to do two things.

1. To represent the choice data in a space of two binary attributes: Social $\in$ {Selfish, Generous} and Risk $\in$ {Safe, Risky} (Section A.3.3). We construct the space so that an observer who believes Exley's null hypothesis would expect the decision maker to be close to indifferent in all choices, i.e. so *ambivalence* is satisfied (see Section 5).

2. To impute some choices that are not observed in the data (Section A.3.4).

The first property 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 $\$yL$ for Charity and $\$y'L$ for Self, for $L \geq 0$.

Linearity in payoffs 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 say that the participant values a given lottery more in dollars to Self than in dollars to Charity, she needs to be able to rank certainty equivalents measured in these units. Linearity in payoffs allows her to do so.

The second assumption is that preferences over bundles are preserved under linear rescaling of probabilities, so we refer to it as *separability in probabilities*. 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 $p \in [0, 1]$ (since all lotteries have exactly one non-zero prize, the assumption does not require *linearity* in probabilities).

A.3.3 Constructing a binary attribute space with "ambivalence"

We need to transform the data for two reasons. First, 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: we cannot expect the participant to be close to indifferent. Second, Prize and Probability are multivalued, and so do not immediately fit into a binary attribute representation.

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. An observer who believes in Exley's null hypothesis believes that the decision maker's utility has the following form (ignoring background wealth):

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

Linearity in payoffs is captured by λ. The participant is indifferent between 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.

To these, we add Constant Relative Risk Aversion (CRRA): $v(y) = y^\alpha$, which gives us utility function (5).

$$v(c, y, p) = \pi(p)\left(\frac{y}{1 + \lambda c}\right)^\alpha. \tag{5}$$

Our approach amounts to selecting choices from the choice lists that can be described by two binary attributes that plausibly satisfy ambivalence: Social $\in$ {Selfish, Generous}, and Risk $\in$ {Safe, Risky}. We do the following:

First, we analyze preferences within a set of choice lists defined by a given lottery probability P. We cannot make comparisons across different Ps, because we would not expect ambivalence to be satisfied and because in any case such choices are not observed. Thus, we 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 ambivalence plausibly holds. In essence we ensure that an observer who believed the participant maximizes (5) would expect them to be close to indifferent.

Consider the self lottery $(0, 10, P)$ that pays \$10 to Self with probability P. Equation

(5) implies the following utilities are equal:

$$v \underbrace{(0, 10, P)}_{\text{Self lottery}} = v \underbrace{(1, (1+\lambda)10, P)}_{\text{Charity lottery}} = v \underbrace{\left(0, \pi(P)^{\frac{1}{\alpha}}10, 1\right)}_{\text{Self sure payoff}} = v \underbrace{\left(1, (1+\lambda)\pi(P)^{\frac{1}{\alpha}}10, 1\right)}_{\text{Charity sure payoff}} \quad (6)$$
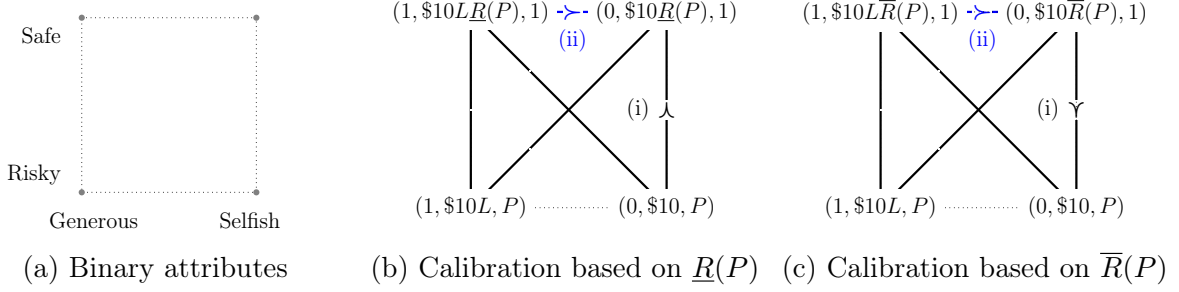
Our approach will be to focus on choices defined by two participant-specific scaling parameters, L and $R(P)$, such that Charity prizes are an L-multiple of self prizes, and sure prizes are an $R(P)$-multiple of risky prizes. So, our binary attribute space consists of: (1) the Self lottery paying \$10 with probability P, (2) the Charity lottery paying \$10L with probability P, (3) the Self sure payment of \$10$R(P)$, and (4) the Charity sure payment of \$10$LR(P)$. 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 own revealed preferences. L is set using the initial normalization choice in the experiment: $L := X/10$ (which is also the rate at which Exley compares self and charity payoffs). 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 therefore implies that the participant can be expected to to have a slight preference for a payoff LY to charity over Y to self, but is also close to indifferent.

We consider two possible values for $R(P)$, set 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 $\overline{Y}(P^S)$. 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 5a. Choice sets calibrated based on $\underline{R}(P)$ allow us to observe cycles in which (Selfish, Risky) is chosen over (Selfish, Safe) (Figure 5b). Choice sets calibrated based on $\underline{R}(P)$ allow us to observe cycles in which (Selfish, Safe) is chosen over (Selfish, Risky) (Figure 5c). We assume that both preferences are close enough to indifferent that ambivalence holds.

Note that Exley's analysis uses the midpoints between just-rejected and just-accepted payoffs to approximate certainty equivalents (i.e. points of indifference) of different lotteries, which are the outcomes in her regression analyses. Our analysis uses the observed choices only, so is expressed in terms of strict preferences.

The CRRA assumption implies that (the log of) (5) can be written as a separable function of the binary attributes Social and Risk, weighted by $\ln\left(\frac{1+\lambda}{L}\right)$ and $\ln\left(\frac{\pi(P)^{\frac{1}{\alpha}}}{R(P)}\right)$ respectively. A derivation is available on request.

(a) Binary attributes   (b) Calibration based on $\underline{R}(P)$ (c) Calibration based on $\overline{R}(P)$

Panels (a) shows four bundles in our binary-attribute representation. Panels (b) and (c) display these bundles in terms of $(Recipient, Prize, Probability)$ where $Recipient = 1$ denotes Charity and $Recipient = 0$ denotes Self. Choice sets marked in black 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*.

Figure 5: Binary attribute representation of Exley (2016)'s choice data

A.3.4 Imputing non-observed choices

Figures 5b and 5c include preferences on the upper horizontal choice set (labeled (ii)), which are not observed in the data. But the observed calibration choice ($\$X$ to charity is preferred to $\$10$ to self), plus *linearity in payoffs* implies that $\$LY$ to charity is preferred to $\$Y$ to self, for all $Y \geq 0$. We use this to impute the ranking of the two Safe bundles.[41]

Our calibration of the binary attribute space is constrained by the lotteries that we observe, whose prizes Exley also calibrated from X, that is, charity lotteries pay $X = 10L$ and self lotteries pay 10. Thus we cannot examine payoffs that vary in other proportions, and therefore cannot observe or impute a choice set where (Selfish, Safe) $\succ$ (Charity, Safe).

A.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.

Null hypothesis 1: homogeneity. The null hypothesis of our first permutation test is that, conditional on P, the probability we observe a given cycle is the same for all participants, i.e. that C_{iP} is iid conditional on P. E.g., it could be that when $P = 0.5$, all

---

[41]We could in principle also use *separability in probabilities* to impute preferences on the lower horizontal choice set. Then, we could construct a figure 8 based on the two observed diagonal choices and two imputed horizontals. But this seems a stronger assumption since we never observe a direct choices between lotteries to self and to charity. Additionally, it would mean we infer implicit preferences from only two decisions.

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 thus asks whether permuting indices i within each probability P reproduces the same distribution over C_{iP}. Intuitively, this test asks whether variation in cycling behavior could be explained by a single representative agent.

Null hypothesis 2: homogeneity conditional on a cycle. The null hypothesis of our second permutation test is that, conditional on P, and *conditional on a cycle being observed* the probability we observe a given cycle is the same for all participants, i.e. that C_{iP} is iid conditional on P and conditional on $C_{iP} \neq 0$. E.g., it could be that when for all participants, $P = 0.5$, 15% of *cycles* are pro-Risky, 30% are pro-Safe, and 55% are pro-Selfish, but some participants are more likely to cycle than others. Our test 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.

Basic testing approach:

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 $(0, 0, 0)$. (Thus the dataset is represented as a distribution over the simplex $\{c \in \{0, ..., 7\}^3 : c_1 + c_2 + c_3 \leq 7\}$, since at most one cycle can be observed per P).

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 6. We strongly reject Null hypothesis 1 ($p < .001$). The main contributor to this rejection seems to be 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 6 percent of participants should exhibit zero cycles across all seven probabilities. We also find (slightly weaker) evidence against Null hypothesis 2 ($p = .065$). Overall we conclude that there is evidence of both systematic and heterogeneous implicit preferences in the sample.
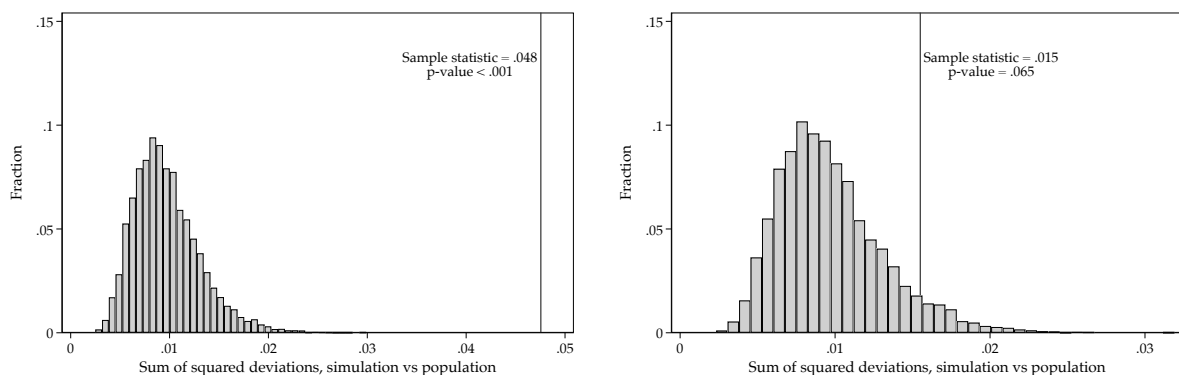

Figure 6: Permutation tests on Exley (2016) data

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

A.3.6 Other analyses

Next, we examine the relationship between implicit preference types. Since each participant has seven opportunities (seven values of P) to reveal their implicit preferences, they could reveal an implicit social preference in some and an implicit risk preference in others. Table A1 shows the full type classification. Overall we find no significant relationship between implicit selfish and implicit risk preferences ($p = .445$).[42]

---

[42]A slight complication is that participants cannot exhibit more than one type of cycle for each P, so a participant with strong pro-Selfish preferences may be less likely to also reveal their implicit risk preference. We see some evidence of this – implicit pro-Selfish participants are slightly more likely to have "Unknown" implicit risk preferences, and less likely to be inconsistent on the risk dimension (for which it is necessary to observe at least two risk cycles). We perform an analysis that attempts to correct for the mechanical relationship: for a participant exhibiting three pro-Selfish cycles there are four remaining "opportunities" to

71

Table A1: Full classification of implicit preference types in Exley (2016)

| Implicit Social | Implicit Risk Preference | | | | |
Preference	Unknown	pro-Safe	pro-Risky	Inconsistent	Total
Unknown	17	13	6	6	42
pro-Selfish	22	13	7	2	44
Total	39	26	13	8	86

Notes: Pearson's $\chi^2(3) = 2.67, p = .445$.

Finally, we examine the relationship between our implicit preference classification and a small number of descriptives measured separately to Exley's main experimental task. We examine (1) whether the "charity" recipient was the American Red Cross (versus another participant in the experiment, see footnote 25), (2) whether the participant exploited "moral wiggle room" in a task modeled on Dana et al. (2007), and (3) participant female gender. Exley analyzes these variables by testing whether participants with these characteristics have larger differences between their cross-context lottery valuations, on average.

Table A2: Predicting descriptives using implicit preference types in Exley (2016)

	(1) Recipient is ARC	(2) Wiggler	(3) Female
pro-Selfish	-0.128	0.172*	0.144
	(0.103)	(0.0934)	(0.110)
pro-Safe	0.00463	0.229*	0.112
	(0.123)	(0.118)	(0.128)
pro-Risky	-0.0289	0.107	-0.125
	(0.159)	(0.142)	(0.151)
Inconsistent	0.194	-0.0261	0.234
	(0.148)	(0.141)	(0.202)
Constant	0.713***	0.108	0.354***
	(0.0919)	(0.0674)	(0.100)
Observations	86	86	86

Robust standard errors in parentheses. $*p < .1, ** p < .05, *** p < .01$. Column (1) outcome equals one if the charity recipient is the American Red Cross (versus another experimental participant). Column (2) outcome equals one if the participant responded to Moral Wiggle Room. Column (3) outcome equals one for female participants.

In Table A2 we regress each of these indicators on our type classification variables. Sim-

---

reveal an implicit risk preference. We regress the number of pro-Selfish cycles on the fraction of remaining opportunities that are pro-Safe and the fraction that are pro-Risky. Neither coefficient is significant and the p-value of an F-test is 0.400. Results available on request.

ilar to Exley, we find some evidence (not significant) that participants are less implicitly selfish when the recipient is charity, versus when the recipient is another participant in the experiment. Also similar to Exley's findings, implicitly selfish participants are significantly more likely to exploit moral wiggle room. Interestingly, those with an implicit pro-Safe preference are also significantly more likely to be do so, which may relate to the fact that the task involves avoiding resolution of uncertainty about the outcome of a potentially selfish choice. Finally we find no significant relationships with gender, though women are slightly more likely to be classified as implicitly selfish and implicitly risk averse.

A.4   Appendix material for analysis of DeSante (2013)

Data access: We accessed DeSante's replication data through the Harvard Dataverse:

DeSante, Christopher, 2013, "Replication data for: Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor", `https://doi.org/10.7910/DVN/AZTWDW`, Harvard Dataverse, V2, UNF:5:EEexoDfcqPKwaPVr7DS6Ow== [fileUNF]

A.4.1   Regression analysis

Equation (2) tells us that under a linearity assumption we can learn about average implicit preferences by comparing average evaluations between comparisons. We cannot identify $v(.)$ directly, but for a change in comparison from $(\boldsymbol{x}, \boldsymbol{z})$ to $(\boldsymbol{x}', \boldsymbol{z}')$ we can identify $\overline{\kappa}_i (\theta_i' - \theta_i)$.

In the case of our analysis of DeSante (2013) we observe evaluations of Black and White candidates with varying work ethic and varying comparators. When race is shared, the influence of race is high (θ_{race}^H) and the influence of work ethic is low (θ_{ethic}^L). When race is non-shared, the influence of race is low (θ_{race}^L) and the influence of work ethic is high (θ_{ethic}^H). We encode attributes such that $White = 1, Good = 1$ (and ignore the "children" attribute).

When work ethic is concealed, we observe the following mean evaluations:

$$\overline{y\left((Black),(Black)\right)} = \overline{v\left((Black)\right)} - \overline{\kappa}_{race}\theta_{race}^H$$

$$\overline{y\left((Black),(White)\right)} = \overline{v\left((Black)\right)} - \overline{\kappa}_{race}\theta_{race}^L$$

$$\overline{y\left((White),(Black)\right)} = \overline{v\left((White)\right)} + \overline{\kappa}_{race}\theta_{race}^L$$

$$\overline{y\left((White),(White)\right)} = \overline{v\left((White)\right)} + \overline{\kappa}_{race}\theta_{race}^H$$

From these we can identify $\overline{v\left((Black)\right)} - \overline{\kappa}_{race}\theta_{race}^L$, $\overline{v\left((White)\right)} + \overline{\kappa}_{race}\theta_{race}^L$, and $\overline{\kappa}_{race}\left(\theta_{race}^H - \theta_{race}^L\right)$. Similarly, when work ethic is revealed we can identify various terms including $\overline{\kappa}_{ethic}\left(\theta_{ethic}^H - \theta_{ethic}^L\right)$.

In the tables we report estimates of $2 \times \overline{\kappa}_{race}\left(\theta_{race}^H - \theta_{race}^L\right)$ and $2 \times \overline{\kappa}_{ethic}\left(\theta_{ethic}^H - \theta_{ethic}^L\right)$, because these capture how the total gap between White and Black (or between Good and

Bad ethic) changes as a result of the change in comparison.

Table 2 in the paper shows only the main coefficients of interest. For completeness we report all coefficients here, in Table A3.

Table A3: Quantitative estimates using Desante (2013) data

	Work ethic concealed	Work ethic revealed
$\overline{v\big((Black)\big)} - \overline{\kappa}_{race}\theta_{race}^L$	594.6	
	(13.12)	
$\overline{v\big((White)\big)} + \overline{\kappa}_{race}\theta_{race}^L$	550.6	
	(12.54)	
$\overline{v\big(\big(\begin{smallmatrix}Black\\Bad\end{smallmatrix}\big)\big)} - \overline{\kappa}_{race}\theta_{race}^L - \overline{\kappa}_{ethic}\theta_{ethic}^L$		475.9
		(23.69)
$\overline{v\big(\big(\begin{smallmatrix}Black\\Good\end{smallmatrix}\big)\big)} - \overline{\kappa}_{race}\theta_{race}^L + \overline{\kappa}_{ethic}\theta_{ethic}^L$		641.0
		(22.92)
$\overline{v\big(\big(\begin{smallmatrix}White\\Bad\end{smallmatrix}\big)\big)} + \overline{\kappa}_{race}\theta_{race}^L - \overline{\kappa}_{ethic}\theta_{ethic}^L$		486.6
		(23.01)
$\overline{v\big(\big(\begin{smallmatrix}White\\Good\end{smallmatrix}\big)\big)} + \overline{\kappa}_{race}\theta_{race}^L + \overline{\kappa}_{ethic}\theta_{ethic}^L$		656.8
		(21.71)
$2 \times \overline{\kappa}_{race}\left(\theta_{race}^H - \theta_{race}^L\right)$	71.13	66.16
	(30.21)	(38.10)
$2 \times \overline{\kappa}_{ethic}\left(\theta_{ethic}^H - \theta_{ethic}^L\right)$		-47.27
		(35.33)
Observations	756	750
Participants	378	375
R-squared	0.861	0.821

Attribute polarities encoded such that $(White, Good) = (1,1)$. Standard errors clustered by participant. Note that $\theta_{race}^H, \theta_{race}^L$ presumably differ between concealed/revealed work ethic.