Implicit Preferences[*]

Tom Cunningham[†] Jonathan de Quidt[‡]

April 25, 2024

Abstract

Simple decisions can reveal two layers of preference. Suppose a hiring manager always chooses a woman over an identically-qualified man, but always chooses a man over a woman with different qualifications. Intuitively, these choices reveal an *explicit* preference for women, but an *implicit* preference for men. We define an implicit preference for an attribute as one whose influence increases whenever that attribute is mixed with others ("dilution"). We prove two representation theorems, and provide three two-layer decision-making models that exhibit implicit preferences. We give extensive guidance for applications, and present evidence of implicit risk preferences, implicit selfishness, and implicit discrimination.

JEL codes: D91, J71, C90

[*] We thank for comments, among others, Ingvild Almås, Roland Benabou, Colin Camerer, Ed Glaeser, Karin Hederos, Marc Kaufmann, Mats Köster, Sendhil Mullainathan, Pietro Ortoleva, Antonio Rangel, Alex Rees-Jones, Anna Sandberg Trolle-Lindgren, Sebastian Schweighofer-Kodritsch, Tomasz Strzalecki, Adam Szeidl, Florian Zimmermann, numerous conference and seminar audiences. We are grateful to Marissa Lepper, Alistair Wilson, and Lise Vesterlund (and their coauthors) for their generosity in sharing data. We thank three anonymous referees for many helpful comments and suggestions. Both authors acknowledge financial support from Jan Wallanders och Tom Hedelius Stiftelse samt Tore Browaldhs Stiftelse, grants I2012-0119:1 (Cunningham), and B2014-0460:1 and BF17-0003 (de Quidt). Previously circulated as "Implicit Preferences Inferred from Choice."

[†] OpenAI. tom.cunningham@gmail.com.

[‡] Queen Mary University of London and Institute for International Economic Studies, Stockholm University. CAGE, CEPR, CESifo, ThReD. j.dequidt@qmul.ac.uk.

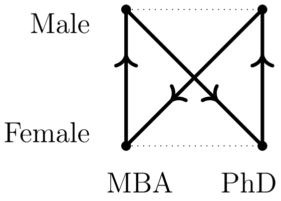
> *"However we may conceal our passions under the veil, there is always some place where they peep out"* - La Rochefoucauld.

Inconsistencies in decision making are often described as arising from a conflict between opposing motives. 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 within pairs of job applicants, a woman and a man, each of whom has 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 study preferences over bundles of binary attributes (Male/Female, Black/White, Aisle/Window). The utility of consuming a given bundle has an *explicit* component that is independent of context, and an *implicit* component, whose strength of influence varies with context. Our core assumption is *dilution*: the influence of an implicit preference for an attribute increases in comparisons that mix that attribute with more other attributes, in a superset sense. Thus our hiring manager explicitly prefers women, but implicitly prefers men. The diagonal choice sets mix gender with qualification, increasing the influence of their implicit gender preference, causing the intransitivity.

The example covers binary choice between bundles, but our theory also applies to *evaluations* such as consumers' willingness to pay, teachers' grading, or judges' sentencing, when each evaluation invokes a comparison. Suppose the manager is setting wages for a pair of new hires, one male and the other female. Our theory says that their implicit preference for men will make the *man's* wage sensitive to the *woman's* attributes. So we would predict that a male candidate would tend to be assigned a lower wage when he is compared to a woman with the same qualification, than when compared to a woman with a different qualification.[1]

Section 1 presents our main theoretical results, which are a pair of representation theorems. Theorem 1 is an application of a Theorem of the Alternative: a representation exists if and only if there is no weighted combination of (1) utility inequalities, from observed behavior; and (2) inequalities on the influence of implicit preferences, from the theory, such that the coefficients on each term sum to zero.[2] Theorem 2 shows that existence of this weighted combination is equivalent to existence of a matching between choices or evaluations. The core intuition is this: we can rule out that the manager implicitly prefers men over women if observed choices include an intransitive cycle, within which every instance where a man is chosen over a woman can be matched to an instance where a woman is chosen over a man *in a more dilute choice set*.[3]

The theorems are stated in abstract terms so that they will be applicable to a broad range of data. Section 2 provides a set of "canonical" examples, e.g. showing that our "figure 8" example reveals an implicit preference favoring men. We describe several other intuitive patterns in choice ("right triangle," "parallel triangles," "square") and in evaluation ("scissor," "parallel scissor"), each of which has distinct implications about implicit preferences.

We claimed above that our theory captures a common intuition. Section 3 demonstrates this by introducing three simple decision-making models, or "foundations," inspired by three separate literatures. In each model, decision makers can possess two distinct attitudes towards each attribute. Each model is nested by our more general model, meaning the two layers of attitude can be identified using our proposed tests.

First we consider a "*ceteris paribus*" decision maker who is subject to a set of rules that apply in "all else equal" situations. This model relates to models in which the decision maker chooses from a subset of elements that are maximal by some other set of rankings (e.g. Manzini and Mariotti (2007, 2012); Masatlioglu et al. (2012); Cherepanov et al. (2013); Ridout (2021)). In this model the figure 8 cycle described above reveals that the hiring manager prefers men, but faces a rule demanding choosing a woman over an equally-qualified man. Diluting the gender attribute disables the rule.

[1] Some of our evaluation applications use an additional assumption, *dominance from attribute k*. It says that an implicit preference has more influence when its attribute is mixed with a special attribute "k," which we usually think of as capturing everything the bundles have in common. Then, implicit gender preferences have more influence when comparing two men than when comparing a man to a woman.

[2] As a referee suggests, this theorem is related to the "cancellation" condition that characterizes additive utility functions (Fishburn, 1970; Wakker, 1989).

[3] The full statement of the theorem is more abstract. It (1) allows for implicit preferences on arbitrarily many attributes, and behavioral data containing multiple cycles; (2) applies to choice and evaluation; and (3) holds for a general class of assumptions about the influence of implicit preferences.

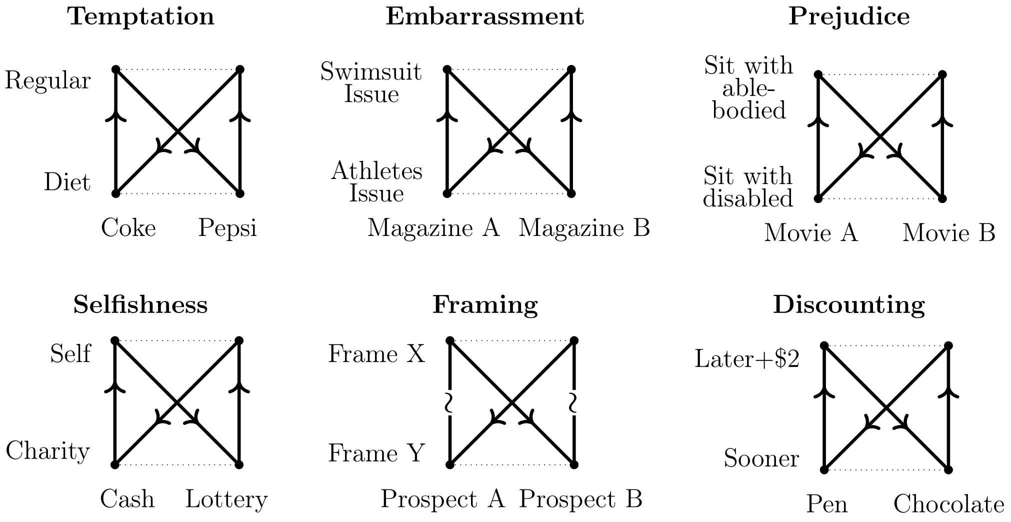
Second, we study a *signaling* decision maker who has intrinsic preferences over each attribute but also cares about others' perceptions of those preferences. This model relates to work on signaling and self-signaling, excuse-driven behavior, and "moral wiggle room" (e.g. Bodner and Prelec (2003); Benabou and Tirole (2003, 2006); Norton et al. (2004); Dana et al. (2006, 2007); Andreoni and Bernheim (2009); Exley (2016a); Bursztyn et al. (2023)). The figure 8 reveals a sincere preference favoring men, but a signaling motive to favor women. Diluting gender adds noise to the observer's inferences, weakening the signaling motive.

Third, we model 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 composed of two rational agents, each with private information. Agent 1, the pre-conscious brain, associates men with high value; it has a "good feeling" about the male candidate. Agent 2, the conscious brain, believes gender-based associations are irrelevant and tries to overrule their influence. Diluting the gender attribute makes it harder for the conscious brain to diagnose and adjust for the source of the "good feeling."

Next we turn to applications. Section 4 gives guidance for using the theory in practice, combined with web Appendix A.2 that provides a "cookbook" of easy-to-use tests. Section 5 applies the theory to three pre-existing datasets. We find evidence of implicit selfishness and implicit risk attitudes in choice data from Exley (2016a) and Ahumada et al. (2022), and implicit racial bias in evaluation data from DeSante (2013a). A simple structural estimation suggests that implicit racial bias shifts average financial allocations by at least 12%.

Finally Section 6 discusses related theories, and explains where our predictions differ from theirs, as well as where there is overlap.

In psychology the term "implicit" applies to attitudes, cognition, judgments, preferences, or knowledge "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 "System 1." Meanwhile, explicit attitudes are those that are stated or revealed deliberately. Psychologists have developed numerous non-choice techniques for identifying implicit attitudes, most notably the Implicit Association Test (IAT) (Greenwald et al., 1998), which is based on variation in response time between different stimuli. IATs have been widely adopted, including within economics (Rooth, 2010; Reuben et al., 2014; Glover et al., 2017; Alesina et al., 2018; Carlana, 2019; Corno et al., 2022), but their interpretation remains controversial (Oswald et al., 2013; Greenwald et al., 2015). A weakness of operationalizing implicit preferences with non-choice data is that their economic significance is unclear. In contrast our definition of implicit preferences is behavioral, so the link to economic outcomes is immediate.



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 or a cash amount, each benefiting themself or charity. They explicitly prefer to give to charity, but reveal implicit selfishness. (Inspired by Exley (2016a)).
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. (This summarizes the findings in Cubitt et al. (2018)).

Figure 1: Figure 8 intransitivities applied to various domains.

Numerous empirical studies share the intuition that underlying motives are revealed when comparisons vary in directness or transparency; we formalize this intuition and provide broadly-applicable tests. Most study discrimination: on disability (Snyder et al., 1979); gender (Norton et al., 2004; Uhlmann and Cohen, 2005; Bohnet et al., 2016); race (Hodson et al., 2002); and bodyweight (Caruso et al., 2009). Exley (2016a) studies excuse-driven selfishness, while Cubitt et al. (2018) study patience. The most closely related paper is Barron et al. (2022)'s lab hiring experiment on gender discrimination. Participants chose between candidates who varied in gender and a skill certification. They find evidence of implicit gender bias, applying a "right triangle" test based on our theory.

Our introductory example shows how we can identify implicit gender discrimination, a topic of great recent interest.[4] But there are many other contexts in which we might expect two layers of preference. Figure 1 shows six different figure 8 cycles across a range of domains, illustrating implicit preferences we might expect to observe.

1 Model

Overview. Preferences will exist over *bundles* of n binary attributes: $\boldsymbol{x} \in \mathcal{X} = \{-1, 1\}^n$, e.g. male/female, aisle/window, sugar/sweetener, sooner/later, risky/safe.[5,6] As a working example we will let attribute 1 be qualification, with $x_1 = -1$ for MBAs and $x_1 = +1$ for PhDs, and attribute 2 be gender, with $x_2 = -1$ for Female and $x_2 = +1$ for Male.

The utility function will be comparison dependent, taking the form $u(\boldsymbol{x}, \boldsymbol{z})$, where $\boldsymbol{x}$ is the bundle being consumed (the *target*) and $\boldsymbol{z}$ is a second *comparator* bundle, to which $\boldsymbol{x}$ is being compared. We express observable data on decision making as inequalities, e.g. $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}', \boldsymbol{z}')$. This formalism allows us to represent data from both binary choice and joint evaluation decisions in a common framework.[7]

We assume a decision maker's *implicit preference* for attribute i is either negative, neutral, or positive, $\kappa_i \in \{-1, 0, +1\}$. Thus $\kappa_2 = +1$ denotes an implicit preference favoring men (equivalently, disfavoring women). We assume that the *influence* of implicit preferences depends on the relationship between target and comparator, specifically on the sets of attributes that are shared and non-shared between $\boldsymbol{x}$ and $\boldsymbol{z}$. These are encoded in the vector $\boldsymbol{\delta} \equiv |\boldsymbol{x} - \boldsymbol{z}|$, which we call the *comparison.* Thus if $\boldsymbol{x} = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ and $\boldsymbol{z} = \left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]$ then $\boldsymbol{\delta} = \left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$. We sometimes talk about the *status* of an attribute in a given comparison $\boldsymbol{\delta}$, this refers to whether it is shared ($\delta_i = 0$) or non-shared ($\delta_i = 2$).

[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 distinguishes between taste-based (Becker, 1957) and statistical (Phelps, 1972; Arrow, 1973) discrimination (possibly inaccurate: Bohren et al. (2023)). Either type can be implicit. Bohren et al. (2022) analyze *direct* and *systemic* discrimination. Direct discrimination early in a woman's career contributes to systemic discrimination later, as her resume ends up weaker than an equally-able man's. Implicit discrimination is a form of direct discrimination, and can be amplified by systemic effects.

[5] This follows the Lancaster (1966) "characteristics" approach, in which the analyst must *a priori* map outcomes into some attribute-space. However this is unavoidable if we wish to study preferences over attributes such as gender, educational qualification, or salt content. Our attributes do not need to be intrinsically binary, but should take no more than two values in the dataset.

[6] We write vectors with a bold font, and $\boldsymbol{x}^T$ will refer to the transpose of $\boldsymbol{x}$. Absolute values of vectors will be element-wise: $|\boldsymbol{x}| = \begin{bmatrix}|x_1| & \dots & |x_n|\end{bmatrix}^T$. Inequalities between vectors will be defined as: $\boldsymbol{x} \geq \boldsymbol{z} \Leftrightarrow x_i \geq z_i \ \forall i$; $\boldsymbol{x} > \boldsymbol{z} \Leftrightarrow x_i \geq z_i \ \forall i, \boldsymbol{x} \neq \boldsymbol{z}$; and $\boldsymbol{x} \gg \boldsymbol{z} \Leftrightarrow x_i > z_i \ \forall i = 1$.

[7] Chambers and Echenique (2016) give a textbook discussion of decision theory as analysis of sets of inequalities, however we do not know of other papers that treat both choice and evaluation inequalities in a common framework, as we do here.

We allow the comparison-independent component of utility to be nonseparable, meaning there can be arbitrary patterns of complementarity among attributes. However we require the effects of comparison $\boldsymbol{\delta}$ on utility to be additively separable. We believe this is a natural way to think about implicit preferences, as applying to attributes rather than combinations of attributes, and gives rise to sharp testable predictions. We provide below a necessary and sufficient condition on observable behavior for this separability to hold.

We will make two substantive assumptions on the direction of implicit influences. First, *dilution*: that the implicit influence of attribute i increases when the set of *other* attributes with the same status as i grows, in a superset sense. Thus, implicit preferences over gender are assumed to have more influence in choice between a man and a woman when the man and woman have different qualifications than when they have the same qualifications. Section 3 will show that this dilution property is common to several decision-making models in which there exists an interaction between two "layers" of preferences or information.

Our second assumption, *dominance from attribute* k, ("dominance-k" for short) says that attribute i's implicit influence will always be greater when i has the same status as a specific attribute k. Thus if k is shared, i's influence is greater when i is also shared, and vice versa. This assumption has no testable implications for choice but there are applications in evaluation. Section 3.4 discusses its interpretation in each foundation model.

Both assumptions share a similar intuition: dilution is about mixing attribute i with a larger set of other attributes, while dominance-k is about mixing i with a single important attribute. In both cases the influence of i's implicit preference increases.

We derive two theorems, each is a different way of expressing conditions under which a set of choice or evaluation observations (a *dataset*) is consistent with a given vector of implicit preferences $\boldsymbol{\kappa}$. Theorem 1 is an application of a Theorem of the Alternative to our problem: a representation exists if and only if there is no weighted combination of (1) utility inequalities from the dataset, and (2) influence inequalities from the dilution/dominance-k assumptions, such that the coefficients on each term sum to zero. Theorem 2 shows that consistency can be expressed with a matching condition. Specifically, a representation exists if and only if the dataset does not contain a subset of inequalities which (1) consists of intransitive cycles among target bundles, and (2) for each attribute i and comparison δ there exists a particular 1:1 matching between the inequalities that satisfies the influence assumptions.

The two theorems express the same condition but in different ways. The first is more suited to numerical or algorithmic verification of feasibility, while the second yields crisp and intuitive answers to questions about what implicit preferences are revealed by a given set of choices, and what choices are inconsistent with the model. Section 1.2 shows that our theorems carry over to a variety of extensions of the basic model.

If we are willing to assume that decision maker has implicit preferences over only a single attribute, e.g. only over gender, then the testable predictions become much simpler. However we often wish to deal with decision makers who may have implicit preferences over multiple attributes, so it is necessary to have the full characterization.

In what follows we will operate at a slightly higher plane of abstraction: we define a partial order over comparisons, $\sqsupseteq_i$, called *influence-dominance*, which ranks comparisons according to the strength of influence of attribute i's implicit preference.We prove our theorems for any influence-dominance relation. Dilution and dominance-k correspond to specific assumptions on $\{\sqsupseteq_i\}_{i=1}^n$. Section 1.2 discusses how we can incorporate extensions to the model, via alternative or additional assumptions on influence dominance.

Setup. A **bundle** is a vector of n binary attributes, $\boldsymbol{x} \in \mathcal{X} = \{-1, 1\}^n$. A **comparative utility function**, $u(\boldsymbol{x}, \boldsymbol{z})$, is a function $u : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We call the first argument $\boldsymbol{x}$ the "target" and the second argument $\boldsymbol{z}$ the "comparator".

We encode all data on decision making as a set of inequalities:

Definition 1 (Dataset). *A **dataset** D is a set of m 4-tuples, $(\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j)_{j=1}^m$, with $\boldsymbol{x}^j, \boldsymbol{z}^j, \boldsymbol{x}'^j, \boldsymbol{z}'^j \in \mathcal{X}$.*

Definition 2 (Consistency). *A dataset D is consistent with a comparative utility function $u : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if, for every $j \in \{1, \ldots, m\}$:*

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

This formalization lets us capture two different types of comparative decision making: binary choice and joint evaluation. Under binary choice a strict revealed preference $\boldsymbol{x} \succ \boldsymbol{z}$ reveals $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{z}, \boldsymbol{x})$, a weak preference reveals a weak inequality, $u(\boldsymbol{x}, \boldsymbol{z}) \geq u(\boldsymbol{z}, \boldsymbol{x})$, and indifference reveals two weak inequalities $u(\boldsymbol{x}, \boldsymbol{z}) \geq u(\boldsymbol{z}, \boldsymbol{x})$ and $u(\boldsymbol{x}, \boldsymbol{z}) \leq u(\boldsymbol{z}, \boldsymbol{x})$.[8]

Joint evaluation describes situations in which the decision maker states an evaluation, e.g., willingness to pay, for each of two bundles presented side-by-side. We write the evaluations as $y(\boldsymbol{x}, \boldsymbol{z})$ and $y(\boldsymbol{z}, \boldsymbol{x})$. We assume that the evaluation y is a strictly increasing function of utility, $y(\boldsymbol{x}, \boldsymbol{z}) = g(u(\boldsymbol{x}, \boldsymbol{z}))$. Because we do not know the function $g(\cdot)$, testable implications come just from inequalities. Notably some types of inequality can be observed in joint evaluation data, but not choice data, e.g. we can observe that the same target bundle receives different evaluations given different comparators, $u(\boldsymbol{x}, \boldsymbol{z}) \lesseqgtr u(\boldsymbol{x}, \boldsymbol{z}')$. To construct a

[8] For example, the choices $\boldsymbol{x} \succ \boldsymbol{x}' \succ \boldsymbol{x}'' \sim \boldsymbol{x}$ 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 $u(\boldsymbol{x}, \boldsymbol{x}'') \geq u(\boldsymbol{x}'', \boldsymbol{x})$.

dataset from joint evaluation we first rank all observed evaluations, then enter an inequality into the dataset for each pair of consecutive evaluations. A perfect equality between two evaluations can be represented by a pair of weak inequalities.[9]

We will repeatedly use a certain type of weighted subset of the dataset, which we call a "cyclical selection." We show below that a dataset is inconsistent with a comparison-independent utility function $u(\boldsymbol{x})$ if and only if it contains a cyclical selection, and the existence of cyclical selections will appear in our theorems.

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 $\boldsymbol{s} \in \mathbb{N}^m$ over elements of the dataset such that each target bundle appears equally often on the left- and right-hand sides. I.e., for every $\boldsymbol{x} \in \mathcal{X}$,*

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

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

If the dataset is derived from choice data then, because every target bundle appears equally often on the left- and right-hand sides, a cyclical selection $\boldsymbol{s}$ must be composed of intransitive cycles among the bundles, e.g. $\boldsymbol{x} \succ \boldsymbol{x}' \succeq \boldsymbol{x}'' \succ \boldsymbol{x}$.[10] An example of a cyclical selection from joint evaluation is a single inequality of the form $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}, \boldsymbol{z}')$, i.e. where the same target bundle receives a different evaluation when the comparator changes.

Separable implicit preferences. We now introduce an explicit structure on preferences.

Definition 4. *A* ***separable comparative utility function*** *has the form:*

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

with $v : \mathcal{X} \to \mathbb{R}$, $\kappa_i \in \{-1, 0, 1\}$, $\theta_i : \{0, 2\}^n \to \mathbb{R}_+$, $f : \mathbb{R} \to \mathbb{R}$ and strictly increasing.

[9] For example, the evaluations $y(\boldsymbol{x}, \boldsymbol{x}') = \$310$, $y(\boldsymbol{x}', \boldsymbol{x}) = \$200$, $y(\boldsymbol{x}, \boldsymbol{x}'') = \$200$, $y(\boldsymbol{x}'', \boldsymbol{x}) = \$150$ correspond to a dataset with $m = 4, \bar{m} = 2$: $u(\boldsymbol{x}, \boldsymbol{x}') > u(\boldsymbol{x}', \boldsymbol{x})$; $u(\boldsymbol{x}', \boldsymbol{x}) \geq u(\boldsymbol{x}, \boldsymbol{x}'')$; $u(\boldsymbol{x}, \boldsymbol{x}'') \geq u(\boldsymbol{x}', \boldsymbol{x})$; $u(\boldsymbol{x}, \boldsymbol{x}'') > u(\boldsymbol{x}'', \boldsymbol{x})$.

[10] By Euler's theorem on directed multigraphs (Jungnickel (2005), p27): if each vertex has the same indegree and outdegree then the edges can be decomposed into directed cycles. Some of these cycles may be composed entirely of weak inequalities, but by our definition at least one must have a strict inequality, and so is an intransitive cycle. Put another way: existence of a cyclical selection is equivalent to violation of the Strong Axiom of Revealed Preferences.

Utility depends on the sum of the explicit value of bundle $\boldsymbol{x}$, and the implicit values from each of the n attributes. Attribute i's contribution to implicit value depends on (1) the decision maker's implicit preference for that attribute, κ_i; and (2) that attribute's *influence*, $\theta_i(|\boldsymbol{x} - \boldsymbol{z}|)$. The outer function f is strictly increasing.

The substantive assumption expressed in this formalism is that implicit preferences affect each attribute separately, e.g. decision makers can have implicit preferences over gender and implicit preferences over race but they do not interact. Put another way, we allow for arbitrary patterns of complementarity between attributes, but we assume that changes to the comparison will not change that complementarity: if attribute i and j are complements under $\boldsymbol{\delta}$ then implicit preferences cannot cause them to be substitutes under $\boldsymbol{\delta}'$.

Proposition 1 gives an exact statement of conditions for (1) to be consistent with a dataset. We will show below that this proposition follows as a corollary of Theorem 1.

Proposition 1 (Cancellation)**.** *A dataset D is consistent with some separable comparative utility function if and only if there exists no cyclical selection $\boldsymbol{s}$ such that for every $i \in \{1, \ldots, n\}$ and $\boldsymbol{\delta} \in \{0, 2\}^n$,*

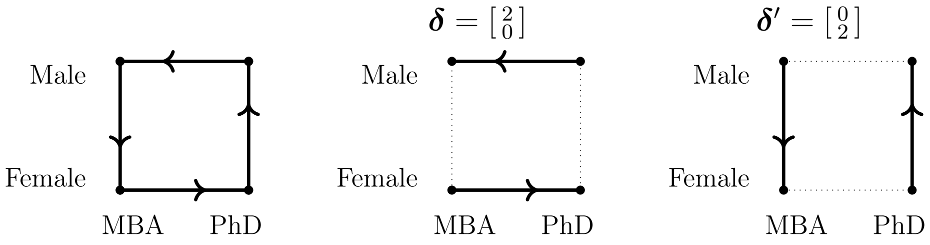
$$\underbrace{\sum_{j:|\boldsymbol{x}^j - \boldsymbol{z}^j| = \boldsymbol{\delta}} s_j x_i^j}_{\textit{inequalities with } \boldsymbol{\delta}^j = \boldsymbol{\delta}} = \underbrace{\sum_{j:|\boldsymbol{x}'^j - \boldsymbol{z}'^j| = \boldsymbol{\delta}} s_j x_i'^j}_{\textit{inequalities with } \boldsymbol{\delta}'^j = \boldsymbol{\delta}} .$$

Thus separability will be falsified if the dataset contains a cyclical selection (an intransitive cycle or set of cycles), in which for each $(i, \boldsymbol{\delta})$ pair, every left-hand side appearance of $x_i = 1$ is "canceled" by a right-hand side appearance of $x_i = 1$ or a left-hand side appearance of $x_i = -1$ (and vice versa). The condition can be seen as a version of the standard "cancellation" condition for separability of a utility function over a discrete multiattribute space (Fishburn, 1970).[11] That condition says that each realization of each attribute must appear "equally often to the left of the preferences ... as to the right of the preferences." (Wakker, 1989). Our condition differs in (a) additionally requiring that the selection constitutes a cyclical selection; (b) aggregating cancellations over attribute-comparison pairs $(i, \boldsymbol{\delta})$ instead of over attributes (i), and (c) incorporating symmetry between positive and negative attribute values.

In the 2-attribute case our assumption rules out the "square" cycle shown in Figure 2. It can be seen that our cancellation condition is violated because: (a) the choices form a cycle; (b) for each $\boldsymbol{\delta}$ and i each realization of i is equally-often preferred and dispreferred.

In this cycle the complementarity between gender and qualification reverses: under $\boldsymbol{\delta}$

[11] We are grateful to a referee for pointing out this connection to us.


$\boldsymbol{\delta} = \left[\begin{smallmatrix}2\\0\end{smallmatrix}\right]$
$\boldsymbol{\delta}' = \left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$
Male
Female
MBA
PhD
Male
Female
MBA
PhD
Male
Female
MBA
PhD

(a) Square cycle (b) Individual comparisons

Figure 2: Example dataset ("square cycle") that violates the condition of Proposition 1.

Males and PhDs are complements, while under $\boldsymbol{\delta}'$ they are substitutes.[12] This reversal of complementarity is ruled out by our separable utility function, and we believe this restriction is a desirable property of an attribute-wise theory of implicit preferences: implicit preferences for attributes should change the relative value of different attributes, but not their patterns of complementarity. In addition a theory which allowed for changes in complementarity would have fewer testable predictions, e.g. it would not rule out the square cycle and it is unclear whether there are any patterns of choice which could falsify such a theory.

Influence dominance. Proposition (1) allows us to test the separability of implicit influences but cannot be used to identify the direction of implicit preferences. For that we need substantive assumptions on how comparisons affect the strength of implicit preferences, i.e. how $\boldsymbol{\delta}$ affects θ_i. We express these assumptions somewhat abstractly, by defining a binary relation over the space of comparisons. The advantage of the abstraction is that our results will be modular, holding for any assumptions on influence that can be expressed in this way.

Definition 5 (Influence-dominance)**.** *An influence-dominance relation is a partial order over the set of comparisons* $\Delta \equiv \{0,2\}^n$.

We will use influence-dominance relations to express assumptions on an influence function $\theta_i(\cdot)$. We will say that θ_i *obeys* influence-dominance relation $\sqsupseteq_i$ if for all $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0,2\}^n$,

$$\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}' \implies \theta_i(\boldsymbol{\delta}) \geq \theta_i(\boldsymbol{\delta}').$$

In words, if comparison $\boldsymbol{\delta}$ influence-dominates $\boldsymbol{\delta}'$ with respect to attribute i ($\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'$), then the influence of i's implicit preference is greater under $\boldsymbol{\delta}$ than $\boldsymbol{\delta}'$ ($\theta_i(\boldsymbol{\delta}) \geq \theta_i(\boldsymbol{\delta}')$).

[12] Abusing notation to write $u(\boldsymbol{x}, \boldsymbol{\delta})$ instead of $u(\boldsymbol{x}, \boldsymbol{z})$, we observe the following:

$$\left(u(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \boldsymbol{\delta}) - u(\left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right], \boldsymbol{\delta})\right) - \left(u(\left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right], \boldsymbol{\delta}) - u(\left[\begin{smallmatrix}-1\\-1\end{smallmatrix}\right], \boldsymbol{\delta})\right) > 0 \quad \text{(complements under } \boldsymbol{\delta}\text{)}$$
$$\left(u(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \boldsymbol{\delta}') - u(\left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right], \boldsymbol{\delta}')\right) - \left(u(\left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right], \boldsymbol{\delta}') - u(\left[\begin{smallmatrix}-1\\-1\end{smallmatrix}\right], \boldsymbol{\delta}')\right) < 0 \quad \text{(substitutes under } \boldsymbol{\delta}'\text{).}$$

Our representation theorems are stated given any arbitrary set of relations $\{\sqsupseteq_i\}_{i=1}^n$. However our applications will use two specific assumptions on each $\sqsupseteq_i$, both based on the "foundations" in Section 3. We present those in Section 1.1 below.

We can now define what it means for a dataset to be consistent with maximization of an implicit preferences utility function:

Definition 6 (Implicit Rationalization by $\boldsymbol{\kappa}$)**.** *We say a dataset D has an implicit rationalization by $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$ if and only if it there exists a rationalization by a separable comparative utility function $u(\cdot, \cdot)$, with implicit preferences $\boldsymbol{\kappa}$, and with $\theta(\cdot)$ obeying the influence-dominance relations, $\{\sqsupseteq_i\}_{i=1}^n$.*

Our first theorem gives a necessary and sufficient condition for implicit rationalization by $\boldsymbol{\kappa}$.

Theorem 1 (Rationalization by Vector)**.** *Given influence-dominance relations $\{\sqsupseteq_i\}_{i=1}^n$, a dataset D has an implicit rationalization by $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$ if and only if there do not exist vectors $\boldsymbol{p} \in \mathbb{N}^m$, with $p_j > 0$ for some $j \leq \bar{m}$, and $\boldsymbol{q} \in \mathbb{N}^{n2^n2^n}$, such that $\forall \boldsymbol{x} \in X$:*

$$\underbrace{\sum_{j:\{\boldsymbol{x}^j = \boldsymbol{x}\}}}_{\substack{\textit{appearances of } \boldsymbol{x} \\ \textit{on LHS}}} p_j = \underbrace{\sum_{j:\{\bar{\boldsymbol{x}}^j = \boldsymbol{x}\}}}_{\substack{\textit{appearances of } \boldsymbol{x} \\ \textit{on RHS}}} p_j,$$

and $\forall i \in \{1, \ldots, n\}, \boldsymbol{\delta} \in \{0, 2\}^n$,

$$\underbrace{\sum_{j:|x^j - z^j| = \boldsymbol{\delta}}}_{\substack{\textit{inequalities with} \\ \boldsymbol{\delta} \textit{ on LHS}}} p_j x_i^j \kappa_i \quad - \underbrace{\sum_{j:|x'^j - z'^j| = \boldsymbol{\delta}}}_{\substack{\textit{inequalities with} \\ \boldsymbol{\delta} \textit{ on RHS}}} p_j x_i'^j \kappa_i \quad + \underbrace{\sum_{\bar{\boldsymbol{\delta}}':\boldsymbol{\delta} \sqsupseteq_i \bar{\boldsymbol{\delta}}'}}_{\substack{\textit{comparisons} \\ \textit{dominated by } \boldsymbol{\delta}}} q_{i\boldsymbol{\delta}\bar{\boldsymbol{\delta}}'} \quad - \underbrace{\sum_{\bar{\boldsymbol{\delta}}:\bar{\boldsymbol{\delta}} \sqsupseteq_i \boldsymbol{\delta}}}_{\substack{\textit{comparisons} \\ \textit{dominating } \boldsymbol{\delta}}} q_{i\bar{\boldsymbol{\delta}}\boldsymbol{\delta}} = 0.$$

The condition comes from writing the assumptions as a system of inequalities, and then applying a Theorem of the Alternative to that system. The first expression is exactly the definition of a cyclical selection: i.e. there must exist some weighted subset of the dataset such that each bundle appears equally-often on the left-hand side and the right-hand side.

The second expression is more complicated to interpret. Loosely speaking it requires that the decision maker consistently chooses bundles with implicitly-preferred attributes in comparisons in which implicit preferences have weaker influence.

We can interpret each of the two pairs of terms in the second condition as "flows." The first pair of terms represents the positive and negative preferences for attribute i under comparison $\boldsymbol{\delta}$ expressed across the cyclical selection $\boldsymbol{p}$, and so their sum represents the excess preference for attribute i under comparison $\boldsymbol{\delta}$, which can be either positive or negative. By itself this is basically a "cancellation" condition (Fishburn, 1970; Wakker, 1989). The second two terms represent inflows and outflows between pairs of comparisons which satisfy influence dominance. For each $\boldsymbol{\delta}$ outflow is to the set of comparisons dominated by $\boldsymbol{\delta}$, the inflow is the comparisons dominated by $\boldsymbol{\delta}$. Thus the condition as a whole says that there exists some "flow" between pairs of comparisons which exactly balance. The next section introduces a matching formalism which we believe is a much more intuitive, but equivalent, condition.

The full proof is in Section 8. The proof first rewrites the dataset and the influence-dominance relations 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. Applying 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, which gives us the two conditions above, i.e. the existence of vectors $\boldsymbol{p}$ and $\boldsymbol{q}$.

Proposition 1 follows directly from Theorem 1, with $\boldsymbol{s}$ substituted for $\boldsymbol{p}$, where we let the set of influence-dominance relations $\{\sqsupseteq_i\}_{i=1}^n$ be empty (equivalently, set $\boldsymbol{q} = \boldsymbol{0}$). Then, the second two terms in the second condition drop out, as do the κ_i's. Intuitively, if $\boldsymbol{q} = \boldsymbol{0}$, implying no restrictions from influence dominance, any $\boldsymbol{\kappa}$ can rationalize D unless the excess preference for every attribute equals exactly zero.

Matching. We next establish that rationalization can be expressed as a matching condition. The matching condition is logically equivalent to the vector condition of Theorem 1 but is easier to interpret and to verify in routine cases. Heuristically, the matching condition verifies that bundles that are implicitly preferred are not systematically ranked *higher* in comparisons in which the influence of implicit preferences is *lower*.[13]

First, we define what is being matched. We define a "score" for each attribute i and each comparison $\boldsymbol{\delta} \in \{0,2\}^n$, which measures how frequently the positive value of that attribute "wins," in the sense that $x_i = 1$ appears on the left-hand side of an inequality or $x_i = -1$ appears on the right-hand side, relative to how many times it "loses" (the converse).

Definition 7 (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** is a vector $\boldsymbol{c} \in \mathbb{Z}^{n\{0,2\}^n}$, with one element for each $i \in \{1, \ldots, n\}$ and each $\boldsymbol{\delta} \in \{0,2\}^n$. Its elements equal:*

$$c_{i,\boldsymbol{\delta}} = \underbrace{\sum_{j:|\boldsymbol{x}^j-\boldsymbol{z}^j|=\boldsymbol{\delta}} s_j x_i^j}_{\textit{inequalities with } \boldsymbol{\delta}^j = \boldsymbol{\delta}} - \underbrace{\sum_{j:|\boldsymbol{x}'^j-\boldsymbol{z}'^j|=\boldsymbol{\delta}} s_j x_i'^j}_{\textit{inequalities with } \boldsymbol{\delta}'^j = \boldsymbol{\delta}}.$$

In a cyclical selection each bundle $\boldsymbol{x} \in \{-1,1\}^n$ necessarily appears equally often on the left- and right-hand sides, so for each i the sum of scores across comparisons is zero: $\sum_{\boldsymbol{\delta}} c_{i,\boldsymbol{\delta}} = 0$.

[13] We are not aware of other representation theorems in multiattribute utility that use matching conditions.

Observe also that $\boldsymbol{c} = \mathbf{0}$ is equivalent to the condition in Proposition 1.

We now define our matching. Roughly speaking, it asks whether increases in influence are associated with more positive scores:

Definition 8 (influence-positive matching)**.** *Given a dataset $D = \{x^j, z^j, x'^j, z'^j\}_{j=1}^m$, a cyclical selection $\boldsymbol{s} \in \mathbb{N}^m$, and an influence-dominance relation $\sqsupseteq_i$, we say attribute i has an* ***influence-positive matching*** *if there exists a matrix of non-negative integers $M_i \in \mathbb{N}^{\{0,2\}^n \times \{0,2\}^n}$ with:*

$$\begin{aligned}
&\forall \boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0,2\}^n, && (M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'} > 0) \implies \boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}' && \textit{(matches obey dominance)}\\
&\forall \bar{\boldsymbol{\delta}} \in \{0,2\}^n, && c_{i,\bar{\boldsymbol{\delta}}} = \underbrace{\sum_{\boldsymbol{\delta} \in \{0,2\}^n} M_{i,\bar{\boldsymbol{\delta}},\boldsymbol{\delta}}}_{\textit{outflow: } \bar{\boldsymbol{\delta}} \textit{ dominates}} - \underbrace{\sum_{\boldsymbol{\delta}' \in \{0,2\}^n} M_{i,\boldsymbol{\delta}',\bar{\boldsymbol{\delta}}}}_{\textit{inflow: } \bar{\boldsymbol{\delta}} \textit{ dominated}} && \textit{(net flows are matched)}
\end{aligned}$$

The first condition says that $\boldsymbol{\delta}$ is only matched to $\boldsymbol{\delta}'$ under attribute i if $\boldsymbol{\delta}$ influence-dominates $\boldsymbol{\delta}'$ with respect to i. The second condition checks that all matchings add up: for every comparison $\bar{\boldsymbol{\delta}}$ the score $c_{i,\bar{\boldsymbol{\delta}}} > 0$ is equal to the net matching flow from other comparisons, i.e. the outflows minus the inflows. So, a positive score (attribute i wins more than it loses for comparison $\boldsymbol{\delta}$) needs to have net outflows ($\boldsymbol{\delta}$ needs to influence-dominate more than it is influence-dominated). A negative score needs to have net inflows.

We likewise say that i has an **influence-negative matching** if there is an M_i that satisfies the first condition (matches obey dominance) and the inverse of the second condition, i.e. outflow minus inflow sums to $-c_{i,\bar{\boldsymbol{\delta}}}$. An attribute with all scores equal to zero trivially has both kinds of matching. We can now state the matching result.

Theorem 2 (Rationalization by matching)**.** *Given influence-dominance relations $\{\sqsupseteq_i\}_{i=1}^n$, a dataset D has an implicit rationalization by $\boldsymbol{\kappa} \in \{-1, 0, 1\}^n$ if and only if there exists no cyclical selection $\boldsymbol{s}$ such that, (1) every attribute with a positive implicit preference ($\kappa_i = 1$) has an influence-negative matching, and (2) every attribute with a negative implicit preference ($\kappa_i = -1$) has an influence-positive matching.*

The full proof is in Section 8. The key point is to show that the existence of the row-weighting vector $\boldsymbol{q}$, from Theorem 1 is equivalent to the existence of a cyclical selection with appropriate influence-negative and influence-positive matchings.

To apply the matching result one must (1) figure out the set of possible cyclical selections;[14] (2) for each $\boldsymbol{s}$, work through each attribute and ask whether it has an influence-

[14] For a single cycle one only needs to consider one, because $\boldsymbol{s}$ and $\lambda\boldsymbol{s}$ (where λ is a positive integer) admit equivalent matchings. When D contains multiple cycles there will be many possible cyclical selections, but typically all empirical content will be contained in just a few of them.

positive, influence-negative, or both kinds of matching (can I match comparisons in which $\boldsymbol{x}$ is preferred, to comparisons in which $\boldsymbol{x}$ is dispreferred, with weakly higher/lower influence?) Section 2 gives many worked examples. For simple datasets this can often be evaluated by visual inspection. In more complex cases the search can sometimes be simplified by using the matrix representation of the problem from Theorem 1.

Theorem 2 illuminates a useful fact that simplifies analysis of choice data: when looking for matchings for attribute i we can ignore choices in which i is shared. This is because $\boldsymbol{\delta}$ is identical on the LHS and RHS of a choice inequality, and shared i implies $x_i = x'_i$. As a result, that choice contributes zero to $c_{i,\boldsymbol{\delta}}$ and will not affect whether it is possible to find an influence-positive or influence-negative matching for i. Another way to see this is by inspection of (1): the terms corresponding to implicit influences from shared attributes will cancel out in any choice inequality $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{z}, \boldsymbol{x})$.

The set of $\boldsymbol{\kappa}$'s that could rationalize the dataset are those not ruled out our Theorems. 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 1/2.*

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

Proof: if $\boldsymbol{\kappa} = \mathbf{0}$ the matching condition is trivially satisfied for any cyclical selection. If a dataset contains no cyclical selection it is rationalizable by any $\boldsymbol{\kappa}$.

1.1 Assumptions on Influence Dominance

We now describe two assumptions on the influence-dominance relations, $\sqsupseteq_i$.

Assumption 1 (Dilution)**.** *For all $i \in \{1, \ldots, n\}$, $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0, 2\}^n$:*

$$\underbrace{(\delta_i = \delta'_i)}_{\substack{i \text{ has same status} \\ \text{in comparisons } \boldsymbol{\delta} \text{ and } \boldsymbol{\delta}'}} \quad \wedge \quad \underbrace{\{j : \delta_j = \delta_i\} \supseteq \{j : \delta'_j = \delta_i\}}_{\substack{\text{a superset of attributes have the same status} \\ \text{as } i \text{ in } \boldsymbol{\delta} \text{ compared to } \boldsymbol{\delta}'}} \quad \Longrightarrow \quad \underbrace{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'}_{\substack{\boldsymbol{\delta} \text{ influence-dominates} \\ \boldsymbol{\delta}' \text{ with respect to } i}}$$

Dilution says that the implicit preference on attribute i has more influence in comparisons where i is more "mixed" with other attributes, in the sense that a superset of other attributes have the same status as i (whether that status is shared or non-shared). For example, an implicit preference favoring men will have a weak influence when $\boldsymbol{x}$ and $\boldsymbol{z}$ differ *only* on gender, becoming stronger as $\boldsymbol{x}$ and $\boldsymbol{z}$ differ on other attributes in addition to gender.[15] Thus in our leading example in the Introduction, implicit gender preferences have more influence in the diagonal choice sets than in the verticals, because the diagonals mix gender with qualification $(\theta_2([\begin{smallmatrix}2\\2\end{smallmatrix}]) \geq \theta_2([\begin{smallmatrix}0\\2\end{smallmatrix}]))$.

Dilution is our most important assumption: it is sufficient for all our results on choice and many results on evaluation. However in some evaluation applications there is a natural second assumption, which we call "Dominance from attribute k" or "Dominance-k" for short. It specifies a special attribute, k, such that attribute i's influence increases when i has the same status as k.[16]

Assumption 2 (Dominance from attribute k)**.** *For all* $i \in \{1, \ldots, n\} \setminus k$, $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0, 2\}^n$,

$$\underbrace{(\delta_i = \delta_k) \wedge (\delta_i' \neq \delta_k')}_{\substack{\textit{i has same status as k in comparison } \boldsymbol{\delta} \\ \textit{i has different status from k in comparison } \boldsymbol{\delta}'}} \qquad \Longrightarrow \qquad \underbrace{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'}_{\substack{\boldsymbol{\delta} \textit{ influence-dominates} \\ \boldsymbol{\delta}' \textit{ with respect to i}}}$$

If k is a shared attribute, influence will be greater for other attributes $i \neq k$ when they are also shared, and vice versa when k is non-shared.

For intuition consider a "signaling" situation where a decision maker wants to conceal their gender bias. k represents an attribute whose importance is very uncertain from the observer's perspective. Then, if i has the same status as k it will be hard for the observer to figure out which is driving the evaluation; k provides cover or an excuse for the decision maker to favor men. Section 3.4 provides conditions under which the assumption holds in each foundation, and further intuition.

For the remainder of the paper we will assume Dilution (Assumption 1 on $\{\sqsupseteq_i\}_{i=1}^n$), and we will indicate when we additionally assume Dominance-k for some k.[17]

Assumptions 1 and 2 can be falsified independently of the separability of the utility function, i.e. we can find datasets consistent with (1) but violating Assumptions 1 or 2, meaning that there is no $\boldsymbol{\kappa}$ that rationalizes the data given the assumptions. Web Appendix

[15] Conversely, when $\boldsymbol{x}$ and $\boldsymbol{z}$ have the same gender (gender is shared), the influence of gender becomes *weaker* as the set of non-shared attributes grows. This part of dilution is not relevant for choice, because implicit preferences on shared attributes do not matter, but it is for evaluation.

[16] Dominance-k has no additional implications on top of Dilution for choice. This is because shared attributes do not affect the set of matchings we can construct in choice datasets: we only need to know what happens to the influence of a non-shared attribute when other attributes change status.

[17] Formally, we let $\sqsupseteq_i$ be the partial order induced by the transitive closure of the union of partial orders induced by the two assumptions. Assumptions 1 and 2 each invoke a partial order over $\{0, 2\}^n$ (each is reflexive, transitive, and antisymmetric). Their transitive closure is trivially reflexive and transitive. To see that it is also antisymmetric, we must show that their union does not include a cycle. This follows from the fact that dominance-k exclusively ranks all comparisons with $\delta_i = \delta_k$ above those with $\delta_i \neq \delta_k$, while dilution never ranks such a pair.

A.1 gives an example. The example is a dataset containing two cycles that under Dilution jointly rule out all $\boldsymbol{\kappa}$'s, but do not violate the cancellation condition of Proposition 1.

1.2 Generalizing the Model

The core matching result in this paper can be generalized to a broader class of context-dependent utility functions, $u(\boldsymbol{x}, \delta)$, where $\delta \in \Delta$ can represents any contextual information about the decision. We can define a generalized separable implicit preferences utility function:

$$u(\boldsymbol{x}, \delta) = f\Big(v(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \kappa_i \theta_i(\delta)\Big).$$

Given a set of influence-dominance relations $\sqsupseteq_i, i \in \{1, \ldots, n\}$ defined over Δ, Theorems 1 and 2 will continue to hold, meaning the testable implications of this extended model can be characterized by an analogous matching condition.

We see four natural ways to use this extended definition of comparison. First, we could generalize the notion of a pairwise comparison, e.g. let $\delta \equiv (\boldsymbol{x} - \boldsymbol{z})$ instead of $\delta \equiv |\boldsymbol{x} - \boldsymbol{z}|$, meaning the *direction* of differences matters (with appropriate generalizations of Assumptions 1 and 2).[18] Second, we could model alternative sources of variation in influence, such as time pressure, the presence of an observer, or whether choices are incentivized. If we believe that implicit preferences have more influence in "fast" decisions, then this implies $\delta \sqsupseteq_i \delta'$ for all i if δ has more time pressure. Third, we could allow for "sequential" comparison effects coming from previously-considered bundles (as in e.g. Kessler, Low, and Shan (2023)). Then, $\boldsymbol{x}$ could represent the current bundle under consideration and δ a comparison with the prior one. Fourth, δ could allow for different-sized comparison sets. For instance, we could model "separate evaluation" as in Hsee et al. (1999). Suppose we set $\boldsymbol{\delta} = \mathbf{0}$ when $\boldsymbol{x}$ is evaluated without a comparator (i.e., all attributes are shared). Assumptions 1 and 2 would then imply implicit preferences have more influence in separate than in joint evaluation.

2 Canonical Examples

We now give a number of important examples of choice and evaluation datasets, and explain what they reveal about the decision maker's implicit preferences. These examples are especially useful for applications, because they yield sharp identification with few decisions. We assume Dilution holds throughout (Assumption 1) and introduce Dominance-k (Assumption

[18] The Signaling-Choice foundation in Section 3 assumes the observer has mean-zero priors over attributes' intrinsic values. If we relax that assumption the direction of attribute differences will matter. We could also allow influence to vary for each choice set ($\delta \equiv \{\boldsymbol{x}, \boldsymbol{z}\}$) or each bundle within a choice set ($\delta \equiv (\boldsymbol{x}, \boldsymbol{z})$).

2) when we get to evaluation. Web Appendix A.2 formally defines each class of example in general terms that allow for arbitrarily many attributes.

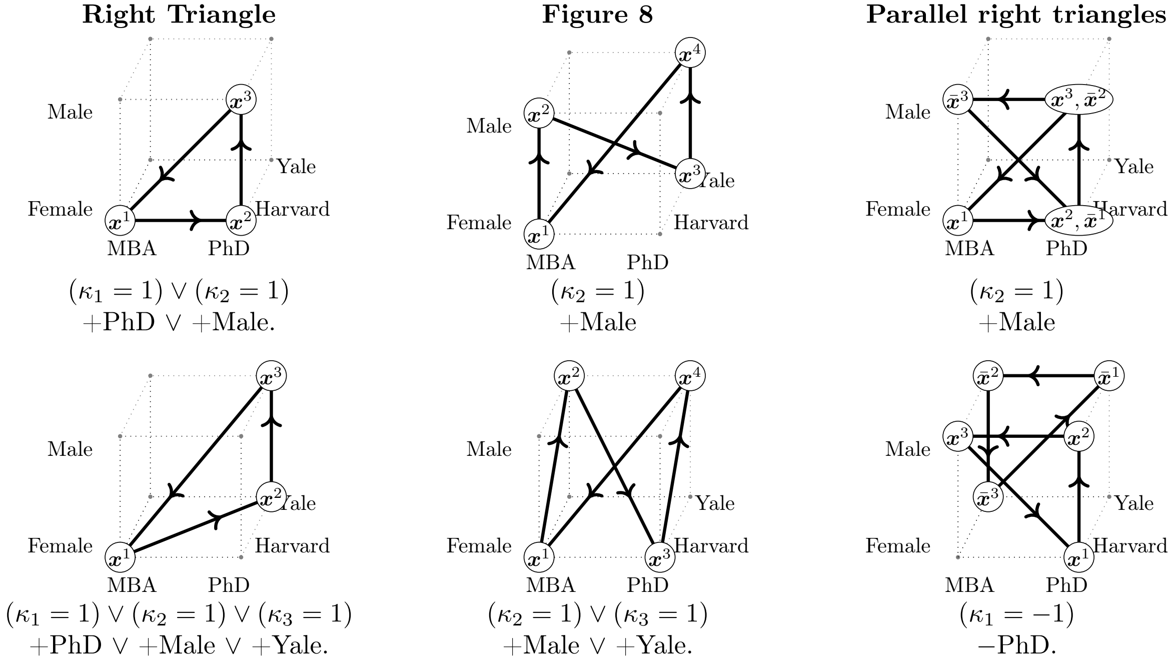
In our examples, attribute 1 is always qualification (PhD = 1), attribute 2 is gender (Male = 1), attribute 3 is college (Yale = 1). We state each example's implications for $\boldsymbol{\kappa}$, and in natural language. In natural language, we always state preferences relative to the +1 pole of the attribute using this shorthand: "+Male" means an implicit preference favoring men (relative to women), "−Male" means an implicit preference favoring women (i.e., against men), and "±Male" means we learn there is an implicit gender preference but not its sign.

Graphical representation of inequalities. We can use intuitive diagrams to visualize inequalities in a dataset. Bundles are represented as points in n-dimensional space. A solid arrow between bundles $\boldsymbol{x} \rightarrow \boldsymbol{x}'$ represents an inequality between two targets: $u(\boldsymbol{x}, \cdot) > u(\boldsymbol{x}', \cdot)$. A dashed line indicates a comparison, so for $u(\boldsymbol{x}, \boldsymbol{z})$ we draw $\boldsymbol{x} \text{ - - - } \boldsymbol{z}$. In the case of choice, where each target is the other's comparator, the dashed and solid lines coincide.

Thus the top-left panel of Figure 3 shows a three-bundle choice cycle $\boldsymbol{x}_1 \succ \boldsymbol{x}_2 \succ \boldsymbol{x}_3 \succ \boldsymbol{x}_1$, while the top-left panel of Figure 4 shows a single inequality from evaluation: $u(\boldsymbol{x}, \boldsymbol{z}_2) > u(\boldsymbol{x}, \boldsymbol{z}_1)$. Both are examples of cycles since they begin and end with the same target.

Choice examples. Take the first right triangle in Figure 3, a three-bundle cycle in which gender and qualification vary. This decision maker has a strictly positive implicit preference on qualification ($\kappa_1 = 1$, favoring PhDs), gender ($\kappa_2 = 1$, favoring men), or both. Suppose not, i.e. suppose both were weakly negative, $(\kappa_1 \leq 0) \wedge (\kappa_2 \leq 0)$, weakly favoring women and MBAs. Informally, we see the decision maker choose an MBA over a PhD (horizontal), and a female over a male (vertical). The diagonal choice set is a dilution of the horizontal with respect to qualification and of the vertical with respect to gender, increasing the influence of both implicit preferences, so they should also choose the Female MBA over the Male PhD. Instead we see the opposite, a contradiction.

Formally, we apply Theorem 2 as follows. The dataset contains a single cycle with three choice inequalities: $u(\boldsymbol{x}^1, \boldsymbol{x}^2) > u(\boldsymbol{x}^2, \boldsymbol{x}^1)$, $u(\boldsymbol{x}^2, \boldsymbol{x}^3) > u(\boldsymbol{x}^3, \boldsymbol{x}^2)$, and $u(\boldsymbol{x}^3, \boldsymbol{x}^1) > u(\boldsymbol{x}^1, \boldsymbol{x}^3)$. Any cyclical selection must put equal weight on each inequality, so we focus without loss of generality on $\boldsymbol{s} = [1, 1, 1]^T$. There are three unique comparisons: $\boldsymbol{\delta}^1 = [2, 0, 0]^T$ (horizontal), $\boldsymbol{\delta}^2 = [0, 2, 0]^T$ (vertical), and $\boldsymbol{\delta}^3 = [2, 2, 0]^T$ (diagonal). The score vectors have four nonzero elements $\boldsymbol{c}_{1,\boldsymbol{\delta}^1} = -2$, $\boldsymbol{c}_{1,\boldsymbol{\delta}^3} = 2$, $\boldsymbol{c}_{2,\boldsymbol{\delta}^2} = -2$, $\boldsymbol{c}_{2,\boldsymbol{\delta}^3} = 2$. Dilution tells us the diagonal is a dilution of the horizontal with respect to attribute 1 (so $\boldsymbol{\delta}^3 \sqsupseteq_1 \boldsymbol{\delta}^1$), and is a dilution of the vertical with respect to attribute 2 (so $\boldsymbol{\delta}^3 \sqsupseteq_2 \boldsymbol{\delta}^2$). Therefore there exists an influence-positive matching for attributes 1 and 2, while attribute 3 trivially has both influence-positive and influence-negative matchings (its score equals zero for all $\boldsymbol{\delta}$ values). This rules out any $\boldsymbol{\kappa}$ with both $\boldsymbol{\kappa}_1 \leq 0$ and $\boldsymbol{\kappa}_2 \leq 0$. We conclude that at least one must be strictly positive, i.e., this decision maker has an implicit preference favoring PhDs, or men, or both.



Right Triangle

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

Figure 8

$(\kappa_2 = 1)$
+Male

Parallel right triangles

$(\kappa_2 = 1)$
+Male

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

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

$(\kappa_1 = -1)$
−PhD.

Figure 3: Examples of choice datasets and their implications

Take now the first Figure 8. Informally, we see a shift of preference from women to men when we dilute the gender attribute by moving from the vertical to the diagonal choice sets. The other attributes are only non-shared in the diagonal choice sets, but because these have the same comparisons (all attributes are non-shared), the influence of implicit preferences must be the same in both. As a result, no attribute other than gender can explain the cycle, so we conclude the decision maker must implicitly favor men.[19]

Consider the first "parallel right triangles." The bottom triangle reveals an implicit preference favoring PhDs and/or men, the top triangle favors MBAs and/or men. The decision maker must have an implicit preference favoring men to rationalize both.[20]

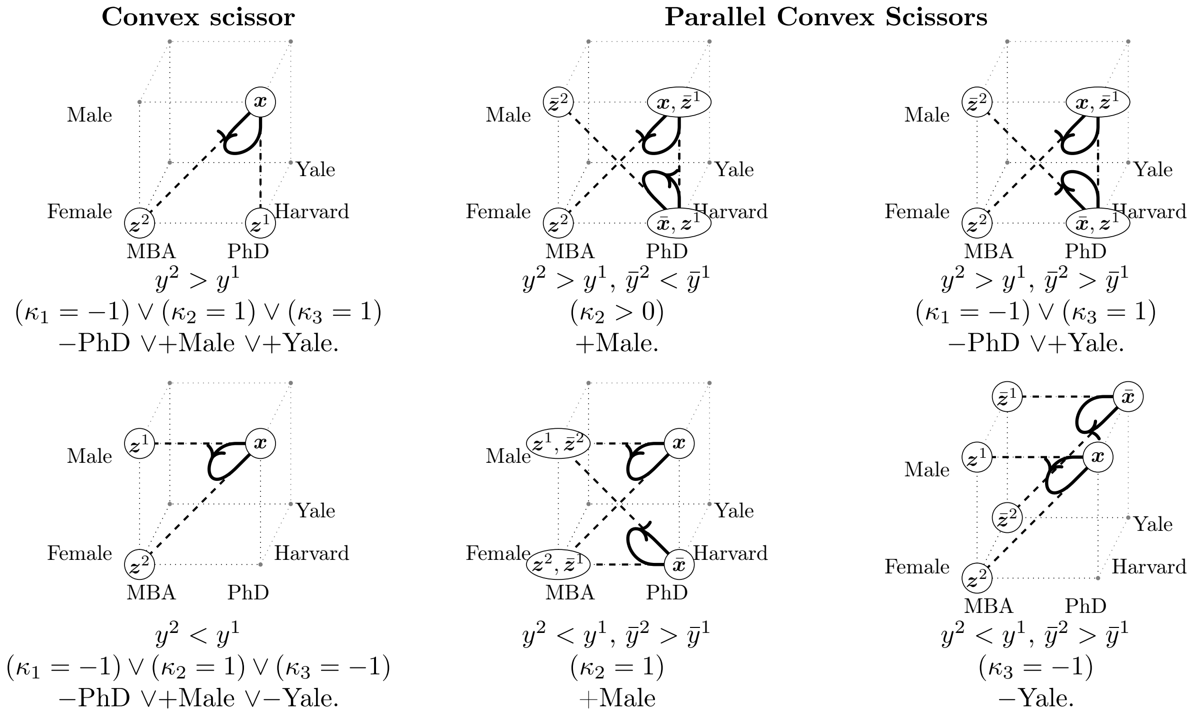
[19] Formally, focus w.l.o.g. on $\boldsymbol{s} = [1,1,1,1]^T$. There are just two unique comparisons: $\boldsymbol{\delta}^1 = [0,2,0]^T$ (vertical) and $\boldsymbol{\delta}^2 = [2,2,2]^T$ (diagonal). The latter is a dilution of the former with respect to the gender attribute. The score vectors have just two nonzero elements $\boldsymbol{c}_{2,\boldsymbol{\delta}^1} = -4$, and $\boldsymbol{c}_{2,\boldsymbol{\delta}^2} = 4$. Thus, attribute 2 (gender) has an influence-positive matching, all other attributes trivially have both kinds of matching. This allows us to rule out any $\boldsymbol{\kappa}$ with $\boldsymbol{\kappa}_2 \leq 0$, so we conclude this decision maker must implicitly favor men.

[20] Formally, we must check *all* possible cyclical selections. Any cyclical selection must consist of s "copies" of the bottom triangle and $\bar{s}$ copies of the top one. The inequality corresponding to the vertical choice set, which appears in both triangles, will have weight $s + \bar{s}$. Our conclusion continues to hold for all non-negative integers $s, \bar{s}$. In particular, when $s = \bar{s}$ (e.g., a cyclical selection that counts each triangle once), we find that the gender attribute has an influence-positive matching, while all other attributes have zero scores.

For the remaining examples we leave it to the reader to verify their conclusions.

Evaluation examples. In choice, we could ignore implicit preferences on shared attributes; we cannot for evaluation data. As a result, an evaluation cycle can have implications for implicit preferences on *every* attribute. We also need to make use of our Dominance-k assumption (Assumption 2) to sign changes in influence for attributes that change status. In this section we will assume that the College attribute (which is always shared in our examples) satisfies the assumption, so $k = 3$. See Web Appendix A.2 for analysis of datasets with arbitrarily many attributes, with and without imposing Assumption 2.

Our examples consist of "convex scissors," which are cycles of the form $u(\boldsymbol{x}, \boldsymbol{z}^1) \gtrless u(\boldsymbol{x}, \boldsymbol{z}^2)$. In a convex scissor the second comparison is always a dilution of the first with respect to one or more attributes. We label the evaluation values $y^1 = y(\boldsymbol{x}, \boldsymbol{z}^1)$ and $y^2 = y(\boldsymbol{x}, \boldsymbol{z}^2)$, and use $\bar{y}^1, \bar{y}^2$ notation for the second scissor in a pair.


Male
Female
Yale
Harvard
MBA
PhD
$\boldsymbol{x}$
$\boldsymbol{z}^1$
$\boldsymbol{z}^2$
$\boldsymbol{x}, \bar{\boldsymbol{z}}^1$
$\bar{\boldsymbol{x}}, \boldsymbol{z}^1$
$\bar{\boldsymbol{z}}^2$
$\boldsymbol{z}^1, \bar{\boldsymbol{z}}^2$
$\boldsymbol{z}^2, \bar{\boldsymbol{z}}^1$
$\bar{\boldsymbol{x}}$
$\bar{\boldsymbol{z}}^1$

Convex scissor

Parallel Convex Scissors

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

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

$y^2 > y^1$, $\bar{y}^2 > \bar{y}^1$
$(\kappa_1 = -1) \vee (\kappa_3 = 1)$
$-$PhD $\vee$+Yale.

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

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

$y^2 < y^1$, $\bar{y}^2 > \bar{y}^1$
$(\kappa_3 = -1)$
$-$Yale.

Figure 4: Examples of evaluation datasets and their implications

Take the first example. $\begin{bmatrix}\text{PhD}\\\text{Male}\\\text{Harvard}\end{bmatrix}$ receives a higher evaluation when compared to $\begin{bmatrix}\text{MBA}\\\text{Female}\\\text{Harvard}\end{bmatrix}$ (diagonal) than when compared to $\begin{bmatrix}\text{PhD}\\\text{Female}\\\text{Harvard}\end{bmatrix}$ (vertical). The diagonal is a dilution of the vertical with respect to gender, while the vertical is a dilution of the diagonal with respect to college. Thus gender has more influence in the diagonal than the vertical, and vice-versa for college. Qualification is shared in the vertical and non-shared in the diagonal. We assumed shared attributes have greater influence (via dominance-k), so qualification has less influence in the diagonal. In sum, the observed increase in evaluation cannot be rationalized by weak implicit preferences favoring PhDs *and* women *and* Harvard. We obtain a disjunction: they must have an implicit preference favoring at least one of MBAs, men, or Yale.

Formally, we have one inequality: $u(\boldsymbol{x}, \boldsymbol{z}^1) < u(\boldsymbol{x}, \boldsymbol{z}^2)$. Assume without loss of generality the cyclical selection puts weight 1 on it. There are two comparisons: $\boldsymbol{\delta}^1 = [0, 2, 0]^T$ (vertical), and $\boldsymbol{\delta}^2 = [2, 2, 0]^T$ (diagonal). The score has six nonzero entries: $c_{1,\boldsymbol{\delta}^1} = -1$, $c_{1,\boldsymbol{\delta}^2} = 1$; $c_{2,\boldsymbol{\delta}^1} = -1$, $c_{2,\boldsymbol{\delta}^2} = 1$; $c_{3,\boldsymbol{\delta}^1} = 1$, $c_{3,\boldsymbol{\delta}^2} = -1$. Dilution tells us that $\boldsymbol{\delta}^2 \sqsupseteq_2 \boldsymbol{\delta}^1$ and $\boldsymbol{\delta}^1 \sqsupseteq_3 \boldsymbol{\delta}^2$, while Dominance-$k$ tells us $\boldsymbol{\delta}^1 \sqsupseteq_1 \boldsymbol{\delta}^2$. Thus, attribute 1 has an influence-negative matching, and attributes 2 and 3 have influence-positive matchings. That implies we can rule out any $\boldsymbol{\kappa}$ which satisfies $(\kappa_1 \geq 0) \wedge (\kappa_2 \leq 0) \wedge (\kappa_3 \leq 0)$, giving us our disjunction.

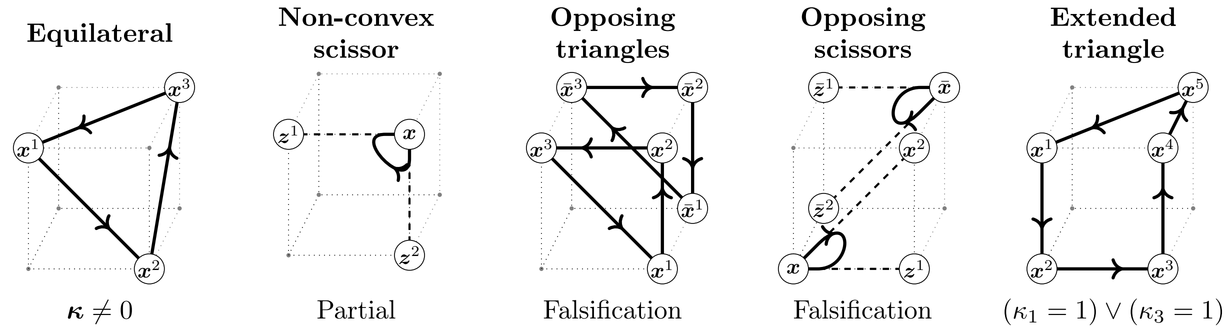
Next we introduce pairs of "parallel convex scissors." These refine identification by eliminating parts of the disjunction identified by a single scissor. In the first example, a male candidate receives a higher evaluation when the influence of gender increases, while a female candidate receives a lower evaluation. In contrast, both candidates have PhDs, from Harvard, yet their evaluations move in opposing directions in response to a change in influence of those attributes. Thus the decision maker must implicitly prefer men, while their implicit preferences on the other attributes are unrestricted.[21] The third example (top-right) is the same but its second scissor has the opposing sign, so it eliminates the gender attribute instead. For the remaining examples we leave it to the reader to verify their conclusions.

Further examples. Figure 5 shows some further examples and their implications.

3 Foundations

We now provide three models of two-layer preferences, each of which can provide a foundation for implicit preferences. In the *ceteris paribus* foundation an implicit preference is a true preference that is constrained by a rule, which applies when certain attributes are shared. Diluting an attribute can cause rules to switch off. In the *signaling* foundation, an implicit preference is a true preference that is sometimes concealed due to a signaling motive. Diluting an attribute weakens the signaling motive. In the *implicit associations* foundation, an implicit preference is an unconscious positive or negative association with an attribute, that

[21] Formally, we have two inequalities, with the same two comparisons as the single scissor. Both bundles have identical values of attributes 1 and 3, but evaluations move in opposite directions when switching from vertical to diagonal. In the cyclical selection that puts equal weight s on both scissors, these attributes' scores equal zero, while attribute 2 has $c_{2,\boldsymbol{\delta}^1} = -2s$ and $c_{2,\boldsymbol{\delta}^2} = 2s$. Thus attribute 2 has an influence-positive matching, while attributes 1 and 3 have both kinds of matching, ruling out every $\boldsymbol{\kappa}$ with $\boldsymbol{\kappa}_2 \leq 0$.



In an **equilateral triangle**, Dilution does not rank influence for any attribute. We can rule out standard preferences (Corollary 2), but nothing else. In a **non-convex scissor** Dilution does not tell us how influence changes. Dominance-k can pin down influence changes for attributes 1 and 2, but we will be inconclusive about κ_3. A pair of **opposing triangles**, and a pair of **opposing scissors** are examples that violate the cancellation condition of Proposition 1 (similar to the square cycle in Figure 2), so no $\boldsymbol{\kappa}$ can rationalize these datasets. The **extended triangle** is a slightly more complex dataset (a five-element cycle with four $\boldsymbol{\delta}$ values) but easy to solve using the matching result (note that the score for the vertical comparisons is zero, leaving us with the equivalent of a right triangle).

Figure 5: Additional examples of datasets and their implications

the conscious brain regards as sometimes informative, sometimes uninformative. Diluting an attribute makes it harder to distinguish informative from uninformative associations.

To keep the discussion concise, for each foundation we provide the setup of the model and state conditions under which the foundation implies an Implicit Preferences utility function and satisfies Dilution (Assumption 1). At the end, we give conditions under which Dominance-k is also satisfied (Assumption 2). See Web Appendix A.3 for all proofs.

It will be useful to have a notation for the set of shared attributes in 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.

3.1 *Ceteris Paribus* Decision Maker

Suppose our hiring manager is subject to a rule on their behavior: they may choose whichever candidate they prefer, except that they are forbidden from hiring a man over an otherwise identical woman. We state a general model of *ceteris paribus* decision makers who are constrained by rules stating 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 (a moral obligation or personal rule) or external (e.g. a bureaucratic rule).

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 an otherwise identical White man, but when choosing between a White woman and a Black man the decision is governed by whichever rule has more force.

Definition 9. *A **ceteris paribus utility function** has the form:*

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

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

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

Proposition 2. *$u^{CP}(\boldsymbol{x}, \boldsymbol{z})$ is an Implicit Preferences utility function satisfying Dilution.*

Applied to choice λ_i represents a bonus or penalty for choosing a bundle with $x_i = 1$ over an alternative with $z_i = -1$, for example when hiring a Black candidate or ordering a low-calorie meal. Inviolable rules have $\lambda = \infty$. 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.

Consider again the manager who truly 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. Thus, the implicit preference has the opposite sign to the penalty λ_i.

3.2 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 updates their beliefs based on the decision maker's actions. The core intuition is the observer updates less about w_i when attribute i is diluted, weakening the signaling motive regarding that attribute.

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*.[22] 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 4.

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.[23]

Definition 10. *A* ***signaling-choice utility function*** *has the form:*

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

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

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 3. *$u^{SC}(\boldsymbol{x}, \boldsymbol{z})$ is an Implicit Preferences utility function satisfying Dilution.*

Consider a hiring manager that prefers men but wants the observer to believe they prefer women. When candidates vary only on gender, the observer updates a lot about the decision maker's gender preferences from the choice. As additional attributes vary between the

[22] If the observer expects the decision maker to prefer one bundle over another, more dilute comparisons can sometimes be *more* informative about an attribute. E.g., while choosing a male PhD over a female MBA is plausibly less informative about gender preferences than choosing a male PhD over a female PhD, choosing a male PhD over a female Nobel prize winner is clearly *more* informative. Thus our choice model does not capture the effect of monetary incentives as discussed by Benabou and Tirole (2006), since the observer would expect that more money is preferred to less.

[23] 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.

bundles the observer updates less about gender, lowering the reputational cost of hiring a man. The implicit preference κ_i has the opposite sign to the signaling motive λ_i: a motive to signal a preference for women manifests as an implicit preference in favor of men.

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

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

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.[24]

Proposition 4. *$u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ is an Implicit Preferences utility function satisfying Dilution.*

[24] 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 prior on $\boldsymbol{\lambda}$, so our solution continues to hold if the observer has incorrect priors, including full naïveté (believing $\boldsymbol{\lambda} = \mathbf{0}$).

The intuition for implicit preferences in this model is very similar to the choice example, except 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.

3.3 Implicit Associations Decision Maker

Finally we describe a decision maker made up of two agents, each with private information about 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 both $\boldsymbol{x}$ 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 second agent's estimate of $\boldsymbol{x}$'s value will be sensitive to the identity of the comparator $\boldsymbol{z}$ because the first agent's estimate of the value of $\boldsymbol{z}$ can be informative about the associations, $\boldsymbol{\lambda}$.

Definition 12. *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

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

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

In equilibrium the sensitivity of utility to x_i, conditional on a comparison $|\boldsymbol{x} - \boldsymbol{z}|$, will be proportional to a weighted average of the adjustments (the π_js) on all attributes which have the same status as i. Thus a dilution of attribute i can either increase or decrease influence depending on the average π of the attributes with which it is mixed. That violates

Assumption 1 which assumes a monotone effect of dilution on influence. Assumption 1 will hold if we assume agent 1's private information is limited to just two attributes. The econometrician does not need to know in advance which two attributes this applies to.

Proposition 5. *If for all $i > 2$ we have $\sigma_i^2 = 0$ then $u^{IA}(\boldsymbol{x}, \boldsymbol{z})$ is an Implicit Preferences utility function satisfying Dilution.*

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 another possible association, 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.[25]

3.4 Conditions under which Dominance-k (Assumption 2) holds

Finally, we give sufficient conditions for Assumption 2, "Dominance-k" to hold in all foundations that apply to evaluation (the assumption is not relevant to choice).

Proposition 6. *The* Ceteris Paribus *decision maker satisfies Assumption 2 if k is not rule-governed ($\lambda_k = 0$). The* Signaling-Evaluation *decision maker satisfies Assumption 2 if the observer's uncertainty about k exceeds all other attributes combined ($\sigma_k^2 \geq \sum_{i \neq k} \sigma_i^2$). The* Implicit Associations *decision maker satisfies Assumption 2 if k is one of the two attributes about which Agent 1 has private information ($k \in \{1, 2\}$, where $\sigma_i^2 = 0, \forall i > 2$).*

In the *ceteris paribus* model, the intuition depends on k's status. Rules do not apply to shared attributes, so if k is shared, having the same status as k disables the rule. If k is non-shared, all rules are disabled (as k is not itself rule-governed) so influence is constant.

The signaling and implicit associations models share similar intuition: k is an attribute with very uncertain weight or value. Then, when another attribute i has the same status as k, it is difficult for an observer to figure out if i, or k, is driving variation in evaluations.

[25] 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.

A signaling decision maker is thus more free to express their true preferences (increase evaluation of men, decrease evaluation of women). For the implicit associations model, the intuition is similar: there is only uncertainty about two attributes, so mixing i with k reduces the conscious brain's ability to learn about the association λ_i, thus increasing its influence.

In most cases it is intuitive to think of k as an attribute that is shared in every observed decision. This is because in most evaluation situations both bundles will have many things in common, many which have quite uncertain value. But to show intuitively how the assumption can in general apply to shared and non-shared attributes, return to the hiring example and suppose the workers could hold one of two professional awards, A or B, where the relative value of these awards is highly uncertain.

Suppose both candidates have award A. When assigning wages to a man and woman, awarding a higher wage to the man cannot be explained by his award, the woman has it too! But when just evaluating two women, the manager could offer them low wages knowing an observer will be unsure if this is just because award A has little value. Implicit gender preferences will have greater influence when gender is shared.

Now suppose that one candidate has award A and the other has award B (so k is non-shared). Now, if the hiring manager awards a low wage to a woman and a high wage to a man, this could be explained by the woman having the worse certificate. But if he gives low wages to two women, that cannot be explained by their awards. Now, implicit gender preferences will have greater influence when gender is non-shared.

4 Guidance for applications

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

Identification "Cookbook." Web Appendix A.2 expresses each canonical example from Section 2 in a general form. The cookbook can be used to easily read off the implications of existing data, as well as to guide new data collection. Additionally, the derivations in that appendix illustrate how to apply the "vector" approach of Theorem 1.

Multivalued Attributes. Some attributes might take multiple values, for example job candidates with three different types of qualification (MBA/PhD/JD) instead of two. In such a case some attribute values could be grouped together, or the dataset partitioned to focus on parts of the attribute space.

***Ambivalence* in Choice Data.** We recommend focusing on choice sets where participants are *likely to be close to indifferent* ("ambivalence"), for two reasons. First, identification relies on observing intransitive choices, which is unlikely when there are large differences in explicit values $v(.)$. Second, in our signaling-choice model dilution will hold only when the observer has a mean-zero prior over the decision maker's preferences for each attribute. Note that neither problem applies to evaluation analysis.

When there are multiple "non-ambivalent" attributes in the dataset we can group them 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 (2016a)'s data demonstrates this in practice. Exley uses choice lists of multiple binary choices between safe and risky payoffs with different recipients. Participants are unlikely to be close to indifferent in some choices, e.g. between a large safe payoff and a lottery with a very low win probability. We construct a binary attribute space by selecting specific choice sets on the choice lists that are deemed close to ambivalent, based on observed behavior in the sample, and classify participants' implicit preferences based on these choice sets. See Appendix A.4 for details.

One could go further and try to test whether an attribute satisfies ambivalence. The signaling logic says the decision maker must believe an observer expects them to be (close to) indifferent to changes in attribute i, i.e. it is the meta-cognition of the decision maker themself that matters. One could therefore elicit participants' beliefs about "expected" or "average" behavior or attitudes with respect to a given attribute, to examine if this is true.

Pre-specifying an attribute space. Sometimes there will be little doubt as to how to define attributes, not always (e.g. when dealing with multivalued or "non-ambivalent" attributes). So, it may be advisable to pre-specify the attribute space before data collection, to reduce degrees of freedom. This is of course a general issue with attribute-based approaches.

Within-subjects Data. The theory assumes we observe a single decision maker's revealed preferences, i.e., "within-subjects" data. A concern with within-subjects data is order effects, whereby later decisions are influenced by earlier ones, e.g. due to a desire for consistency. A common technique to mitigate order effects is to spread decisions over time, intersperse them with "filler" tasks, or in other ways make earlier decisions less salient or memorable.[26] If order effects are a serious concern we might turn to between-subjects data (one decision per participant). This has different implications for choice and evaluation:

Between-subjects Choice Data. Establishing the presence of intransitivities in between-subjects choice data is challenging, because intransitivity is difficult to distinguish from underlying preference heterogeneity (similar to the Condorcet paradox in voting). One remedy is to make the (strong) assumption of homogeneous preferences plus noise, such as in tests

[26] 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).

for "weak stochastic transitivity." Alternatively, one can test for violations of the Triangle inequality (Regenwetter et al., 2011; Müller-Trede et al., 2015). If each participant makes one choice we would need to observe $Pr(a \succ b) + Pr(b \succ c) + Pr(c \succ a) > 2$ to conclude that at least one has intransitive preferences over $\{a, b, c\}$. I.e. the average choice probability must exceed $2/3$.[27] Implicit preferences would need to be strong to satisfy this condition.

Between-subjects Evaluation Data. Our tools for evaluation data carry over well to between-subjects data with some functional form restrictions (we use this in our DeSante (2013a) application). Suppose we observe $t \in \{1, \ldots, T\}$ iid sampled individuals' evaluations of $\boldsymbol{x}$ with comparator $\boldsymbol{z}$. Suppose $v(.)$ and $\boldsymbol{\kappa}$ are heterogeneous across individuals, with population averages $\overline{v(\boldsymbol{x})}$ and $\overline{\boldsymbol{\kappa}}$, and assume for now that $\boldsymbol{\theta}$ is homogeneous. Assume also that 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:

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

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 $\kappa_i = 1$ is more common in the population than $\kappa_i = -1$. We could also allow heterogeneity of the form $\theta_{i,t} = \alpha_{i,t}\theta_i$, then we would identify $sgn(E[\kappa_i\alpha_i])$.

Structural estimation. Our DeSante (2013a) application also structurally estimates equation (2). This is of interest when it is not just the *sign* but also the *magnitude* of implicit preferences that matters. This is relatively straightforward under evaluation but we believe it would also be feasible with choice data in an appropriately-specified empirical discrete choice model. In practice structurally estimating (2) involves estimating one parameter per bundle equal to $\overline{v(\boldsymbol{x})} + \sum_{i=1}^{n} x_i\overline{\kappa}_i\theta_i(\boldsymbol{\delta})$ for some benchmark comparison $\boldsymbol{\delta}$, plus the *changes* in attribute-wise implicit value when $\boldsymbol{\delta}$ changes: $\overline{\kappa}_i\left(\theta_i(\boldsymbol{\delta}) - \theta_i(\boldsymbol{\delta}')\right)$. In general $\overline{v(\boldsymbol{x})}$ is not identified without a normalization $\theta_i(\underline{\boldsymbol{\delta}}^i) = 0$ for some "minimal-influence" comparison $\underline{\boldsymbol{\delta}}^i$. The minimal-influence comparison $\underline{\boldsymbol{\delta}}^i$ will normally depend on i, because when influence is low for one attribute it is usually high for another.

Choice versus evaluation. Section 2 and Web Appendix A.2 provide a set of tools for identifying implicit preferences from binary choice and joint evaluation. Joint evaluation has a number of advantages for applications. An evaluation "scissor" can pick up small comparison effects, whereas in choice we need strong enough implicit preferences, or sufficiently

[27] For four-element cycles the threshold is $3/4$. As an example, the choice proportions in Snyder et al. (1979)'s experiment do not satisfy the criterion, so could be due to heterogeneous, transitive preferences.

well-calibrated choice sets, to find a cycle. Evaluation analysis does not rely on "ambivalence," and is better suited to between-subject data. Finally, it is parsimonious: we can isolate an implicit preference with three joint evaluations (the first parallel convex scissors in Figure 4); the shortest choice cycle that can do this is the figure 8 (four choices).

5 Applications

5.1 Implicit Risk and Social Preferences

Exley (2016a) studies "the use of risk as an excuse not to give." She conducts two experiments in which participants choose between a risky and a sure payment, where for each the beneficiary can be either themselves or a third party.[28] She uses the choices to construct certainty equivalents that value each lottery (to self and to charity) both in terms of money to self and in terms of money to charity. She finds that, on average, 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 self and charity (both payoffs go to self or both to charity), suggesting *implicit selfishness*.

We apply our theory to Exley's dataset (Exley, 2016b), augmented with data from a replication by Ahumada et al. (2022) (henceforth "Ahumada").[29] We exhaustively classify all implicit preferences revealed in the data, as well as behavior inconsistent with the theory.

We have four main findings. First, we confirm the presence of significant implicit selfishness: 50 percent of participants are implicitly selfish. Second, our approach yields new insights in the form of *implicit risk preferences*: 21 percent of participants become more risk averse when risk is diluted, while 9 percent become more risk tolerant. Third, 17 percent of participants reveal both types of implicit preference, and implicit selfishness is associated with implicit risk tolerance.

Fourth, we compare the original and replication samples. Estimating the same regression specification as Exley, Ahumada et al. (2022) find qualitatively similar average behavior, but with attenuated effect sizes and larger p-values (partly due to a smaller sample). However, our individual-level analysis finds a striking congruence between the implicit-preference type distributions in both samples. This strengthens our earlier conclusions, and suggests that our approach is sensitive enough to distinguish heterogeneous preference types even when their influence is harder to detect in average behavior.[30]

[28] In one experiment the third party is the American Red Cross, in the other it is another participant.

[29] They implemented a slightly shorter version of Exley's design, in which the third party is the University of Pittsburgh Medical Center Children's Hospital. We thank the authors for kindly sharing these data.

[30] Exley briefly mentions some individual-level analysis. Her footnote 29 reports that when the win proba-

Data. We need to do a little work to place the data in a binary attribute framework. Here we provide a brief summary, see Web Appendix A.4 for all details.

Each participant-level dataset contains an initial *normalization* choice used to fix the participant's exchange rate between money to self and money to charity in all subsequent decisions. It elicits an amount $\$X$ payable to charity that is just preferred to \$10 to self.[31]

Each subsequent choice is between a safe payoff and a lottery paying a prize with probability $P \in \mathcal{P}$. In Exley, P takes seven values: $\mathcal{P} = \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$; Ahumada restrict to four: $\mathcal{P} = \{0.05, 0.25, 0.75, 0.95\}$; we account for this in our analysis. There are four kinds of choices: (1) charity gets safe vs lottery; (2) self gets safe vs lottery; (3) charity gets safe vs self gets lottery; (4) self gets safe vs charity gets lottery. Self lotteries pay \$10 and charity lotteries pay $\$X$. The participant selects from a choice list the smallest safe amount that they prefer to the lottery.

Participants never choose between different lotteries, so we set up an attribute space with one attribute per value of P, corresponding to the win probability in that choice list. Those attributes are always shared in any observed choice set. For each P, we define four bundles that vary in two attributes: Risk $\in$ {Safe, Risky}, and Social $\in$ {Generous, Selfish}.[32] The four observed choice sets are shown in black in Figure 6. We do not observe, but can impute, a fifth preference: (Generous, Safe) $\succ$ (Selfish, Safe), which we show in blue.[33]

Analysis. Each participant's dataset contains all of their choices across all values of P. For each P there are five possible cycles or combinations of cycles that we can potentially observe, shown in Figure 6. We omit inequalities that are not part of a cycle, since they

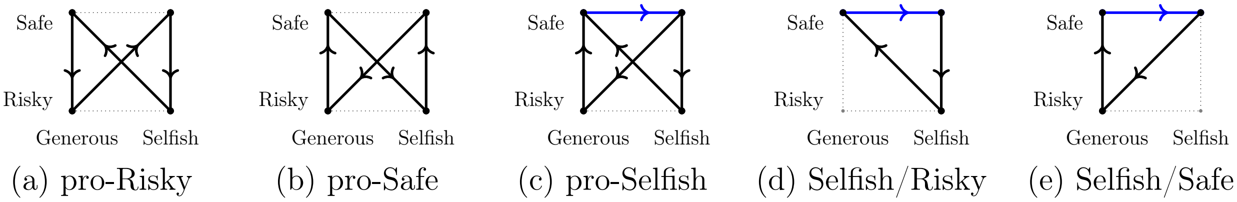
bility is 0.95, 42 percent of participants value the charity lottery strictly lower than the self lottery in terms of money to self, but weakly higher in terms of money to charity. This does not in fact reveal implicit selfishness, but is close to a "pro-Safe" cycle (Figure 6(b)). Our estimates are more conservative because 1) we only count strict preferences (reducing the proportion of cycles when $P = 0.95$ to 26 percent) as true indifferences are not observed; and 2) we focus on "ambivalent" choice sets (reducing it to 15 percent).

[31] Exley mostly excludes participants whose initial normalization choices were censored or inconsistent. We do the same. We pool her two experiments, giving us 86 participants, plus 56 more in Ahumada's data.

[32] The experiment involves choosing between options with different prizes, win probabilities, and recipients. As explained in Section 4, for choice analysis we need "ambivalence" to hold: participants should be "close to indifferent" along each attribute. We would not expect ambivalence to hold between e.g., larger versus smaller prizes, or between identical monetary amounts to self and charity. Instead we construct binary attributes such that Safe bundles pay small prizes for sure (where the size of the prize depends on P, 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 appendix for details.

[33] Participants did not choose between bundles that only differ in the Social attribute. Without these preferences, any pattern in the observed (black) choice sets could be explained by *explicit* selfish preferences. E.g., those in panel (c) can be rationalized by transitive preferences (Selfish, Safe) $\succ$ (Selfish, Risky) $\succ$ (Generous, Risky) $\succ$ (Generous, Safe). 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. Given the data structure, that same assumption implies (Generous, Safe) $\succ$ (Selfish, Safe) (see appendix for details). We never observe any choices corresponding to (Generous, Safe) $\prec$ (Selfish, Safe), so cannot detect any cycles that rely on this preference.

cannot form part of a cyclical selection. Panels (a) and (b) are figure 8 cycles revealing implicit Risk preferences; (c) is two parallel right triangles revealing an implicit Selfish preference; (d) and (e) show right triangles that each reveal a disjunction. Appendix Table A1 shows a data extract for five of Exley's participants, to illustrate what we observe in a typical participant's dataset.



(a) pro-Risky (b) pro-Safe (c) pro-Selfish (d) Selfish/Risky (e) Selfish/Safe

Diagrams show all possible cycles for a given value of P. Choices in black are observable in the data. The blue horizontal choice is imputed. There are a set of additional attributes that encode each value of P. These are always shared so we omit them for simplicity. Implications of each type of cycle are derived in Section 2.

Figure 6: Types of cycle and the implicit preferences they reveal

We can exhaustively classify every observable dataset. Cyclical selections are weighted collections of cycles of type (a), (b), (d), and (e) (since (c) equals one (d) and one (e)), so all empirical content reduces to four indicators counting whether the participant exhibited at least one cycle of the corresponding type.

The simple cases can be read off Figure 6. A participant who never exhibits a cycle is consistent with any $\boldsymbol{\kappa}$ (Corollary 2). If they exhibit one or more (a) or (b) cycles, and nothing else, we identify their implicit Risk preference only. If they exhibit one or more right triangles of the same type ((d) or (e)), and nothing else, we identify a disjunction.

How about a participant that exhibits a combination of different types of cycle? First, any dataset containing at least one (a) and at least one (b) cycle cannot be rationalized by any separable comparative utility function. A cyclical selection that puts equal weight on each will violate the cancellation condition of Proposition 1. Second, any dataset of only (d) and (e) cycles is equivalent to a pair of right triangles, i.e., yields the same set of possible matchings, revealing an implicit Selfish preference, and nothing about Risk.

Any dataset consisting of just type (a) and (d) cycles reveals only a pro-Risky implicit preference, we learn nothing about implicit Selfish preferences. There is no matching we can construct from these observed choices that would also isolate implicit Social preferences or yield a contradiction.[34] If only (b) and (e) cycles are observed we identify a pro-Safe implicit preference and nothing else.

[34] Formally, let a cyclical selection consists of s copies of (a) and s' copies of (d). There are three comparisons: $\boldsymbol{\delta}^1 = [2,0]^T$ (horizontal), $\boldsymbol{\delta}^2 = [0,2]^T$ (vertical), $\boldsymbol{\delta}^3 = [2,2]^T$ (diagonal), where $\boldsymbol{\delta}^3 \sqsupseteq_1 \boldsymbol{\delta}^1$ and $\boldsymbol{\delta}^3 \sqsupseteq_2 \boldsymbol{\delta}^2$. The scores are: $c_{1,\boldsymbol{\delta}^1} = -2s'$, $c_{1,\boldsymbol{\delta}^2} = 0$, $c_{1,\boldsymbol{\delta}^3} = 2s'$; $c_{2,\boldsymbol{\delta}^1} = 0$, $c_{2,\boldsymbol{\delta}^2} = 4s + 2s'$, $c_{2,\boldsymbol{\delta}^3} = -4s - 2s'$. Attribute 2

Next, any dataset consisting of just type (a) and (e) cycles reveals *both* implicit preferences favoring Selfish and Risky, i.e. $\kappa_1 = 1, \kappa_2 = -1$. The latter is an immediate implication of a cyclical selection containing just (a) cycles. To show the former, consider a cyclical selection containing s (a) cycles and $2s$ (e) cycles. We can construct a matching that matches the left-side vertical choice and the SE to NW diagonal choices in (a) to opposing same-$\boldsymbol{\delta}$ choices in (e). The remaining inequalities can be matched in an equivalent way to a pair of pro-Selfish right triangles.[35] Having pinned down both κ's, we can verify by visual inspection that there are no other cyclical selections with contradictory implications. If the dataset were to also contain type (d) cycles the conclusions would not change because there would be no additional matching we could construct that would contradict the prior conclusions. Finally, by equivalent arguments, any combination of (b) and (d) cycles (with or without additional type (e) cycles) reveals implicit preferences favoring Selfish and Safe, i.e. $\kappa_1 = 1, \kappa_2 = 1$.

We apply this classification procedure to Exley's and Ahumada's pooled samples to describe the overall distribution of types. Table 1 presents the fraction of participants corresponding to each preference type. Columns correspond to implicit risk preference types (pro-Risky, pro-Safe, or Unknown). Rows correspond to implicit Social preference types (pro-Selfish or Unknown—the data structure precludes observing implicit Generous preferences). Participants for whom we only learn a disjunction are classified as Unknown.

Implicit preferences are prevalent: 63 percent of participants are classified as having at least one nonzero κ_i. Exactly half of participants are classified as implicitly selfish, while 21 percent are classified as implicitly risk averse and 9 percent as implicitly risk tolerant. Overall, 2.8 percent of participants are inconsistent with the theory.

Table 1 also suggests a correlation between implicit preference types. Implicit risk tolerance is less than half as prevalent as risk aversion, but it is relatively much more common among those who are also implicitly selfish.[36]

Table 2 disaggregates the results of Table 1 in a couple of ways. First, we show type information including cases where we only identify a disjunction. Second, we show results separately for each study. Column 1 reports the full classification for Exley's sample, using all seven values of P. Columns 2 and 3 report the classifications for Exley's and Ahumada's

always has an influence-positive matching in any cyclical selection, while attribute 1 admits influence-positive and -negative matchings, so it must be that $\kappa_2 = 1$, while κ_1 can take any value.

[35] Formally, let the cyclical selection contain 1 "copy" of cycle (a) and 2 copies of cycle (e). Comparisons are the same as in footnote 34. The score vectors have just two nonzero elements $\boldsymbol{c}_{1,\boldsymbol{\delta}^1} = -2$, $\boldsymbol{c}_{1,\boldsymbol{\delta}^3} = 2$. Thus attribute 1 (Social) has an influence-positive matching, while we can construct influence-positive and influence-negative matchings for attribute 2. This rules out any $\boldsymbol{\kappa}$ with $\kappa_1 \leq 0$.

[36] The odds ratio among those with unknown risk tolerance is 0.12 whereas it is 0.85 among the implicitly selfish, a seven-fold increase ($p = .02$, from a logit regression of a dummy for implicit risk aversion on a dummy for implicit selfishness, restricted to those with known risk preference type). We reject the null that risk and social preference types are independent using an exact test ($p = .036$).

Table 1: Frequencies of different types in combined dataset

		Implicit risk preference			
		pro-Risky	Unknown	pro-Safe	Total
Implicit social preference	pro-Selfish	0.077	0.331	0.092	**0.500**
		(0.022)	(0.039)	(0.024)	**(0.042)**
	Unknown	0.014	0.338	0.120	**0.472**
		(0.010)	(0.040)	(0.027)	**(0.042)**
	Total	**0.092**	**0.669**	**0.211**	**0.972**
		(0.024)	**(0.039)**	**(0.034)**	**(0.014)**
	Inconsistent				0.028
					(0.014)

Pooled samples from Exley (2016a) and Ahumada et al. (2022), all observed values of P. Cell entries correspond to the fraction of participants classified as each implicit preference type. Standard errors in parentheses. Row/column-wise totals in bold. If only a disjunction over implicit preferences is revealed we classify the participant as Unknown for both attributes. Inconsistent participants are those exhibiting at least one type (a) and type (b) cycle. $N = 142$ participants. Fisher's exact test for independence of risk and social preference types: $p = 0.036$.

samples, restricting to $P \in \{0.05, 0.25, 0.75, 0.95\}$ to ensure comparability. Column 4 pools both samples and all observed P's (these values match Table 1).

A striking finding in Table 2 is that the type distributions identified using Exley's and Ahumada's samples are quantitatively very similar to one another, and a joint test does not reject equality of the distributions ($p = 0.902$).

Column 5 of Table 2 simulates the expected type distribution under a random choice assumption (each individual preference has a 50 percent chance of pointing in either direction). We do this because some cycles depend on a smaller number of decisions pointing in the "right" direction, so may be observed more often purely due to noise or errors. A joint test strongly rejects equality between the pooled sample type distribution and the prediction from random choice ($p < 0.001$). We also observe a qualitative asymmetry (pro-Safe types are more common than pro-Risky) that is not predicted under random choice. Appendix Table A2 shows the frequencies of each individual cycle type ((a)–(e)) for the same sample restrictions, and compares them to random choice, further reinforcing this conclusion.

Some of the behavioral patterns we observe may be driven by noisy rather than systematic preferences. For example, a participant who is nearly indifferent across several choices might accidentally violate transitivity if they make a mistake when picking from the choice lists. Appendix Table A3 reports a robustness check that requires "stricter" preferences—larger inconsistencies in choice list valuations—before we register a cycle. The number of observed cycles decreases (mechanically), but the overall type classification is quite stable.

In sum, like Exley (2016a), we find substantial inconsistencies in decision making that we

Table 2: Frequencies of different types in Exley & Ahumada et al. datasets

Type		Exley	Exley	Ahumada	Pooled	Random
		7 probabilities	4 probabilities		4/7 probabilities	
Inconsistent	(i)	0.047	0.000	0.000	0.028	0.089
		(0.023)	(.)	(.)	(0.014)	
Pro-Risky only	(ii)	0.012	0.012	0.018	0.014	0.073
		(0.012)	(0.012)	(0.018)	(0.010)	
Pro-Safe only	(iii)	0.128	0.140	0.107	0.120	0.074
		(0.036)	(0.037)	(0.041)	(0.027)	
Pro-Self only	(iv)	0.349	0.314	0.304	0.331	0.255
		(0.051)	(0.050)	(0.061)	(0.039)	
Pro-Risky and Self	(v)	0.093	0.081	0.054	0.077	0.148
		(0.031)	(0.029)	(0.030)	(0.022)	
Pro-Safe and Self	(vi)	0.116	0.093	0.054	0.092	0.148
		(0.035)	(0.031)	(0.030)	(0.024)	
Pro-Risky OR Self	(vii)	0.035	0.047	0.089	0.056	0.082
		(0.020)	(0.023)	(0.038)	(0.019)	
Pro-Safe OR Self	(viii)	0.070	0.151	0.179	0.113	0.082
		(0.027)	(0.039)	(0.051)	(0.027)	
No cycles	(ix)	0.151	0.163	0.196	0.169	0.050
		(0.039)	(0.040)	(0.053)	(0.031)	
Participants		86	86	56	142	

This table shows the classification of participants according to their revealed Preferences in our analysis of data from Exley (2016) and Ahumada et al. (2022). Standard errors in parentheses. First column shows results for all seven values of P in Exley's data. Columns 2 and 3 restrict to $P \in \{.05, .25, .75, .95\}$ for comparability. Column 4 uses all available data, and column 5 simulates random choice for seven and four probabilities respectively in proportion to study sample sizes. **Statistical tests.** Joint test of equality between Exley and Ahumada type distributions (restricted to 4 probabilities): $p = 0.902$. Equality between rows in pooled dataset: $p(ii = iii) < .001$, $p(v = vi) = 0.684$, $p(vii = viii) = 0.101$. Joint versus random choice (pooled dataset): $p < .001$.

attribute to implicit selfishness. We also uncover novel evidence of implicit risk preferences, even though the original study was not designed with them in mind. This highlights the value of a model that allows for implicit preferences on multiple attributes at once. These findings are robust to replication in a new sample and to allowing for noise.

How should we interpret implicit risk preferences? We think it is quite natural to think of them through the lens of the implicit-associations model: the decision maker might consciously be prepared to take a certain amount of risk, but their subconscious or instinctive attitudes could be different (Loewenstein et al., 2001). Then they could be persuaded to take more or less risk by varying the extent to which variation in risk is diluted with other attributes. This could have important implications for real decisions. An implicitly risk-averse decision maker might make more risk-averse choices when choosing between pension plans with different attributes (where risk is more diluted) than when choosing between different variants of the same plan. That could substantially affect long-run wealth.

5.2 Implicit Racial Discrimination

DeSante (2013a) uses a survey experiment to test whether people reward hard work in a "color-blind manner," finding that Black applicants for financial aid are rewarded less for having a good "work ethic" than White applicants (and penalized more for a bad one). He attributes the findings to implicit racism. However, his tests cannot distinguish between implicit and explicit preferences.[37] We reanalyze the data (DeSante, 2013b) using our framework, and find significant evidence of implicit pro-White racial bias, and no significant implicit Work Ethic preference.

In the experiment, participants from a US representative sample made hypothetical state aid decisions for two applicants presented side by side.[38] Applicants vary in Race $\in$ {Black, White}, signaled by their name.[39] In some comparisons Race is shared, in others it is non-shared. Second, in some treatments the applicants' Work Ethic $\in$ {Good, Bad} is revealed, in which case it is always non-shared (in other treatments Work Ethic is concealed).[40] All applicants have two children whose ages are independently randomized. Childrens' ages are not included in the data, so we treat them as an attribute, Children $\in \{c, c'\}$, for which we assume there is no implicit preference. We finally assume a "background" attribute, not shown in our diagrams, that is always shared (e.g., all applicants have in common that they are low-income female parents).

Figure 7 depicts the bundles with concealed Work Ethic that appear in the data (top-left), and those with revealed Work Ethic (bottom-left). Every applicant is evaluated alongside a Black comparator and a White comparator. So, for example, applicant (Black, Bad) is compared with (Black, Good) and with (White, Good).

Dilution (Assumption 1) 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 that the "background" attribute satisfies Assumption 2. Then, the influence of racial preferences is higher when Race is shared than non-shared. This is intuitive: many other shared attributes could "explain" why someone gives a large or small amounts

[37] In some tests he examines the effect of changing the race of the target holding the comparator fixed, or changing both simultaneously. In others, he examines how evaluations change when work ethic is hidden versus revealed, and whether that differs between Black and White applicants. Our analysis shows that to separate implicit from explicit preferences we must hold the target fixed and vary the comparator.

[38] Specifically, they had up to $1,500 to allocate, with any unallocated funds going to "offset the deficit." 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. The budget constraint binds for 31 percent of participants.

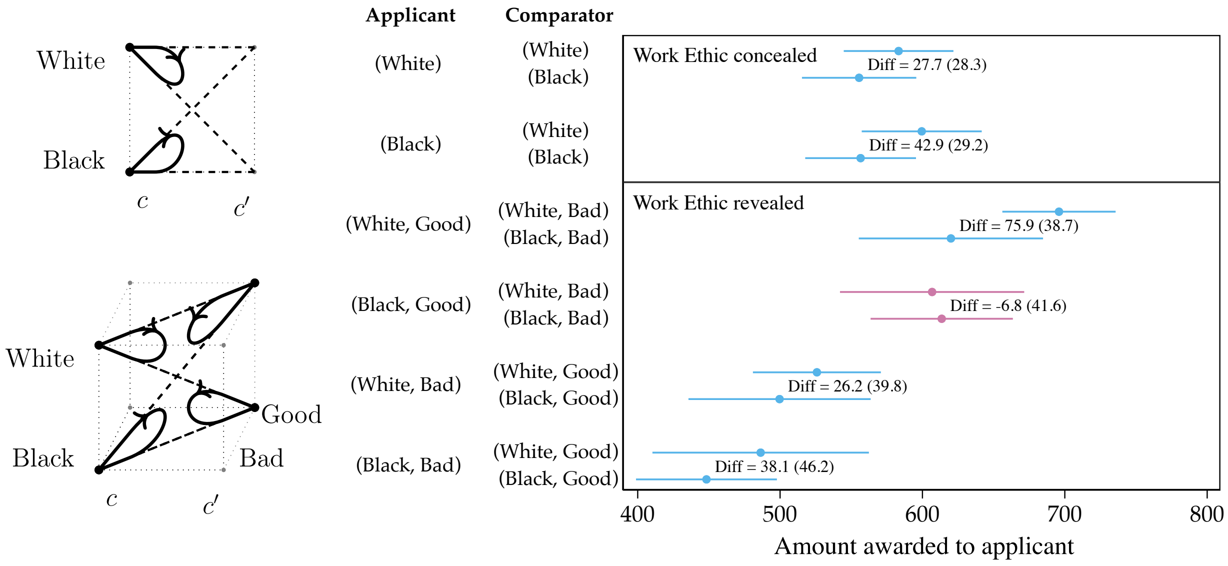
[39] Latoya and Keisha for Black applicants, Laurie and Emily for Whites. Simonsohn (2016) highlights that names can also signal socioeconomic status, so we might be observing implicit preferences over SES.

[40] These are written "Excellent/Poor" in the experiment; we use "Good/Bad" for compactness.

to two applicants with the same Race.

The diagrams show the inequalities we would expect if participants implicitly favor Whites: we expect higher evaluations when the comparator is White than when they are Black. (This matches the bottom-middle example in Figure 4.)

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 4, equation (2)).



Left panel: Attributes are Children (horizontal), Race (vertical), Work Ethic (depth). Always-shared "background" attribute not shown. Variation in "Children" is not observed in the data and assumed irrelevant. Right panel: Pairs of points corresponds to "convex scissors." pro-White implicit preferences predict positive "Diffs" (shown in blue if positive). 95% confidence intervals clustered by participant. $N = 753$ participants.

Figure 7: Reanalysis of DeSante's data

Figure 7 presents the results. We group evaluations in pairs corresponding to six convex scissors. The results suggest pro-White implicit preferences: in all cases but one the evaluation of Blacks decreases, and the evaluation of Whites increase, when Race is shared. The first pair of scissors, when Work Ethic is concealed, also shows a reversal of evaluation: when Race is shared, Whites receive more than otherwise-identical Blacks. When Race is non-shared, Blacks receive more than Whites. That can be interpreted as suggesting that participants *explicitly* favor Blacks, but *implicitly* favor Whites.

The statistical strength of the results is modest. Of the six observed scissors, only one difference is statistically significant. However the mean difference ($34) is highly significant ($p < 0.01$). An F-test of the null that all six differences equal zero has a p-value of 0.08.

As we explain in Section 4 and Web Appendix A.5, our linearity assumption allows us to structurally estimate some parameters, corresponding to implicit preferences on Race, Work Ethic, and the background attribute. By Dilution, the influence of Work Ethic is lowest, and of the background attribute highest, when race is shared. By Dominance-k, the influence of implicit racial preferences is highest when Race is shared. Denote higher influence values by θ_i^H and lower ones by θ_i^L. We can identify $2 \times \overline{\kappa}_i \left(\theta_i^H - \theta_i^L\right)$, which corresponds to the increase in evaluation of a bundle with $x_i = 1$, relative to one with $x_i = -1$, when influence increases from θ_i^L to θ_i^H. E.g., it measures the widening of the gap between Black versus White, when the influence of Race increases.

Table 3 presents our findings. We estimate parameters separately for the hidden/revealed Work Ethic treatments because we would expect influence of other attributes to differ between these treatments. When Race switches from non-shared to shared the gap between White and Black applicants increases by \$71 ($p = .02$) in the Work Ethic concealed treatment, and by \$67 ($p = .08$) in the Work Ethic revealed treatment. This corresponds to about 12% of the mean allocation. Since we only identify the change $\theta_{race}^H - \theta_{race}^L$, that implies that implicit racial preferences can explain at least this share of the total allocation. The coefficients on Work Ethic and the background attribute are smaller and not significant. The Work Ethic coefficient is $-\$47$ ($p = .18$), consistent with an implicit preference against Good applicants (Good applicants get more money in same-race comparisons).

Table 3: Quantitative estimates using DeSante's data

	Work Ethic concealed	Work Ethic revealed
$2 \times \overline{\kappa}_{race} \left(\theta_{race}^H - \theta_{race}^L\right)$	70.67	66.72
	(30.14)	(38.00)
$2 \times \overline{\kappa}_{ethic} \left(\theta_{ethic}^H - \theta_{ethic}^L\right)$		-47.22
		(35.36)
$2 \times \overline{\kappa}_{background} \left(\theta_{background}^H - \theta_{background}^L\right)$	-15.21	35.23
	(48.98)	(47.31)
Participants	378	375
Mean amount awarded	571.8	563.7

$2 \times \overline{\kappa}_{race} \left(\theta_{race}^H - \theta_{race}^L\right)$ equals the change in evaluation of White applicants, relative to Blacks, when influence increases from θ_{race}^L to θ_{race}^H. Second row corresponds to Good relative to Bad applicants, and third row to an always-shared "background" attribute. Standard errors clustered by participant in parentheses.

Overall, the data point to implicit pro-White preferences with economically significant influence. This implies a role for choice architecture: racially biased decisions can be mitigated if decision makers evaluate White and Black applicants simultaneously. However, the statistical strength of the results is modest. A larger sample would give more power, and allow us to test for consistency of implicit preferences across decisions.

6 Relationship to other theories

As we have shown, the model implies substantive restrictions on the data such that we can separately test a number of assumptions: (i) the separability of implicit preferences; (ii) the dilution/dominance-k assumptions; (iii) existence of an implicit preference on a single attribute; (iv) a combination of implicit preferences across multiple attributes.

We now show that these predictions are qualitatively different from existing models in the choice literature. It is worth noting a difference in domain: we consider outcomes with binary attributes, while most existing theory applies either to atomic outcomes or outcomes with scalar attributes.[41] Existing models can be applied to our domain but we will see that they make qualitatively different predictions. We focus particularly on the figure 8 cycle and show that existing models either (1) are inconsistent with a figure-8 cycle; or (2) are consistent but without ruling out other patterns, such as the square cycle. Our empirical choice application finds that numerous participants exhibit figure-8 cycles.

Contingent weighting. There are many theories of multiattribute choice in which the weights on each attribute vary with the choice set, sometimes called "contingent weighting" (Kőszegi and Szeidl, 2012; Cunningham, 2013; Bordalo et al., 2013; Bushong et al., 2020). The models cited are all inconsistent with a figure-8 cycle because the weights on each attribute depend only on the marginal distribution of realizations of that attribute in the choice set, whereas in our model the weight depends on the *joint* distribution of attributes, via the dilution assumption.[42]

Inattention. Inattention models (Sims, 2003; Caplin and Martin, 2014; Woodford, 2012) can rationalize some anomalies in choice. We might expect that choices in which more attributes differ might be more complex, implying dilution would increase complexity, and in

[41] Additionally, our formalism restricts attention to binary comparisons, so we cannot model inconsistencies that stem from adding elements to the comparison set, such as decoy and compromise effects. We discuss of extensions to nonbinary comparisons in the Conclusion.

[42] In Kőszegi and Szeidl (2012) and 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). Suppose utility is entirely separable in each attribute, meaning 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 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 choice with two attributes (Ellis and Masatlioglu (2021)).

turn might make the decision maker less sensitive to attribute variation. This might generate an intransitivity but it would not generate a strict reversal of the direction of preference as we see in the figure-8. Thus we would expect at most to see movement toward indifference (or a convergence in reported evaluations) as comparisons become more complex.

Evaluability. Theories of "evaluability" (Hsee et al., 1999; Hsee and Zhang, 2010) apply to evaluations and assume that people sometimes become more sensitive to an attribute in joint evaluation than separate evaluation.[43] Evaluability effects could thus generate comparison effects on evaluation, but they would not generate a reversal, as can happen in our model (i.e., if $\boldsymbol{x}$ is evaluated above $\boldsymbol{z}$ in separate evaluation, we would not expect that pattern to reverse in joint evaluation).

Inference. Another mechanism that can generate intransitivities is inference from the choice set (e.g. Wernerfelt (1995)). Inference can rationalize any pattern of choices (including a square cycle), the question is what priors would be needed to justify the rationalization. For the Figure 8 cycle in the Introduction it would require that *intrinsic* and *informational* value of gender are opposite, i.e. the decision maker must believe that women are intrinsically better than men, but that the degrees men get are better than the degrees women get. While this might hold in certain situations, we believe our preference-based explanation will typically be more plausible, especially in cases with familiar attributes about which we would not expect a decision maker to substantially update from the choice set.

Noise. Noisy or stochastic choice, or heterogeneity, can cause intransitivities even when the underlying preferences are transitive. We discuss this issue in relation to analysis of between-subjects data (Section 4), and take some practical steps to address it in our empirical applications in Section 5. However, we do not provide a fully-specified model of stochastic implicit preferences. Allen and Rehbeck (2023) is an important related paper insofar as it constructs a random utility model with preferences over attributes (strictly, a hybrid of atomic and attribute-based preferences, with a taste for randomization), and shows how preferences can be identified. However the implied behavioral restrictions are qualitatively different to ours.[44] In the presence of noise, the data requirements for identification from individual behavior are increased due to the need to observe choice probabilities.

Atomic outcomes. A number of theories have a very similar "two layer" motivation to ours: decision makers select from a choice set the element which maximizes their true

[43] See Cunningham (2013) for a Bayesian rationalization. Kahneman and Frederick (2005) discuss a similar phenomenon, that sensitivity tends to be higher in within-subjects than between-subjects experiments. We can extend our framework to capture separate evaluation, see Section 1.2.

[44] The leading example of behavior ruled out by this theory is where a change to bundle $\boldsymbol{x}$ affects the choice probabilities of other bundles in the same choice set but not that of $\boldsymbol{x}$; this is also ruled out in our theory, but only trivially so because we only consider 2-element choice sets, and does not relate to key hallmarks of our theory such as the figure-8 cycle.

preference, from within the subset which are undominated relative to some other relation (or set of relations). The models differ in the nature of the relations: Manzini and Mariotti (2012) assume a single semiorder, Cherepanov et al. (2013) assume multiple binary relations, and Ridout (2021) assumes multiple complete orders. 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 our approach is that implicit preferences can be identified from binary choices.[45] 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 other female candidates as well.

Cubitt et al. (2018). The only paper we are aware of that identifies a figure-8 pattern in choice (besides Cunningham (2016), on which our implicit associations foundation is based) is Cubitt et al. (2018). They outline a model in which sensitivity to an attribute decreases when more attributes vary, but the utility of money does not change. Thus, people are more willing to accept financially-compensated delay when considering two different goods at different time horizons than two identical goods (see Figure 1). More generally, dilution will always increase the relative value of bundles containing money, and that model cannot generate strict cycles over non-monetary attributes, as used in most of our examples.

7 Conclusion

Our paper is motivated by an assumption that is latent in a number of prior studies: people sometimes hold two opposite preferences regarding an attribute and that one preference—the *implicit* preference—has greater influence when the comparison mixes its attribute with others. We formalize that assumption and fully characterize its testable implications.

Our framework is compatible with implicit preferences' influence depending on other contextual factors such as time pressure, cognitive load, stated versus revealed preferences, etc. An advantage of our dilution-based approach is that none of this auxiliary information is required for identification. To the extent that other factors also predict variation in influence we would expect to identify the same sets of implicit preferences. It would also be useful to extend the theory to deal with data on choice proportions, e.g. the fraction of people who choose $\boldsymbol{x}$ over $\boldsymbol{z}$, either assuming an unobserved distribution of deterministic preferences, or irreducibly stochastic preferences as in Allen and Rehbeck (2023).

Some further extensions of interest include generalizing the representation theorem to

[45] 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.

allow for multivalued attributes, or for bundles that are missing some attributes. Suppose our hiring manager chooses a female over a male, a male over hiring nobody, and “no hire” over a female. If an attribute has greater influence when one bundle is missing that attribute (“no hire” is missing gender), this would reveal a pro-male implicit preference.

It is natural to ask how implicit preferences will be revealed in comparison sets larger than two elements. Here, the predictions of our three foundational models diverge. The concept of “influence” is most naturally interpreted as a property of a *choice set* in the implicit associations model, but as a property of a *choice* in the signaling and ceteris-paribus models. The distinction does not matter for binary choice sets, but does for larger sets. Thus our method of identifying implicit preferences from binary choice or joint evaluation necessarily loses some generality when extended in this way.

The theory is relevant for choice architecture design. It predicts that implicit preferences—such as racial or gender bias—have less influence in less-dilute comparisons. So minority-group candidates could be positioned for comparison with majority-group candidates who are similar on other dimensions. See e.g. Bohnet et al. (2016) for related discussion.

We see rich scope for empirical applications, through data reanalysis and fresh experiments, to map out the existence, strength of influence, consistency, and out-of-sample predictiveness of implicit preferences across 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, other kinds of prejudice, 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 two prospects with identical payoffs and different frames, but frames influence choice between non-identical prospects. These, and other applications, we leave to future research.

8 Appendix: Proof of Theorems 1 and 2

To maintain readability in the proofs we use $\boldsymbol{\delta}$ to represent a generic comparison $|\boldsymbol{x}-\boldsymbol{z}| \in \{0,2\}^n$, allowing us to write, e.g., $M_{i,|\boldsymbol{x}-\boldsymbol{z}|,|\boldsymbol{x}'-\boldsymbol{z}'|}$ as $M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'}$.

Proof of Theorem 1

Each inequality in dataset D can written as $v(\boldsymbol{x}^j) + \sum_{i=1}^{n} x_i^j \kappa_i \theta_i(\boldsymbol{\delta}^j) \geq v(\boldsymbol{x}'^j) + \sum_{i=1}^{n} x_i'^j \kappa_i \theta_i(\boldsymbol{\delta}'^j)$, where the inequality is strict for $j \leq \bar{m}$. We can write the two functions, $v(\cdot)$ and $\theta_i(\cdot)$ as vectors $\boldsymbol{v} \in \mathbb{R}^{2^n}$ and $\boldsymbol{\theta} \in \mathbb{R}^{n2^n}$, with elements $v_x = v(\boldsymbol{x})$ (one entry for each $\boldsymbol{x} \in \mathcal{X}$), and $\theta_{i\boldsymbol{\delta}} = \theta_i(\boldsymbol{\delta})$ (one entry for each $i \in \{1, ..., n\}$ and comparison $\boldsymbol{\delta} \in \{0,2\}^n$).

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 for all $i, \boldsymbol{\delta}, \boldsymbol{\delta}'$, $(\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}') \implies (\theta_{i\boldsymbol{\delta}} \geq \theta_{i\boldsymbol{\delta}'})$.

We can write D's inequalities in matrix form with $\left[\begin{smallmatrix} \hat{P} & \hat{X} \end{smallmatrix}\right] \left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right]$ representing the $\bar{m}$ strict inequalities, and $\left[\begin{smallmatrix} \bar{P} & \bar{X} \end{smallmatrix}\right] \left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right]$ representing the $m - \bar{m}$ weak inequalities. Each row corresponds to one inequality from the dataset. The matrix $P = \left[\begin{smallmatrix} \hat{P} \\ \bar{P} \end{smallmatrix}\right] \in \mathbb{Z}^{m \times 2^n}$ holds 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 $X = \left[\begin{smallmatrix} \hat{X} \\ \bar{X} \end{smallmatrix}\right] \in \mathbb{Z}^{m \times n2^n}$ holds coefficients on $\boldsymbol{\theta}$, with entries:

$$X_{\underbrace{j}_{\substack{\text{row} \\ j \in 1,...,m}}, \underbrace{i\boldsymbol{\delta}}_{\substack{\text{column} \\ i \in 1,...,n \\ \boldsymbol{\delta} \in \{0,2\}^n}}} = x_i^j \kappa_i \underbrace{\mathbb{1}\{|\boldsymbol{x}^j - \boldsymbol{z}^j| = \boldsymbol{\delta}\}}_{\substack{=1 \text{ if LHS of inequality } j \\ \text{has comparison } \boldsymbol{\delta}}} - x_i'^j \kappa_i \underbrace{\mathbb{1}\{|\boldsymbol{x}'^j - \boldsymbol{z}'^j| = \boldsymbol{\delta}\}.}_{\substack{=1 \text{ if RHS of inequality } j \\ \text{has comparison } \boldsymbol{\delta}'}}$$

Every inequality in D is satisfied if and only if $\left[\begin{smallmatrix} \hat{P} & \hat{X} \end{smallmatrix}\right] \left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right] \gg 0$ and $\left[\begin{smallmatrix} \bar{P} & \bar{X} \end{smallmatrix}\right] \left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right] \geq 0$.

We encode the influence-dominance relations $\sqsupseteq_i, i = \{1, ..., n\}$ as a matrix of coefficients on $\boldsymbol{\theta}$: $Q \in \mathbb{Z}^{n2^n2^n \times n2^n}$. Q has one row for each combination of an attribute l and pair of comparisons $\bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\delta}}'$ ($n2^n2^n$ rows in total). A row will contain non-zero entries only if $\bar{\boldsymbol{\delta}} \sqsupseteq_l \bar{\boldsymbol{\delta}}'$. If so, the row has entry $+1$ in the column that corresponds to attribute l and comparison $\bar{\boldsymbol{\delta}}$, and has entry -1 in the column corresponding to attribute l and comparison $\bar{\boldsymbol{\delta}}'$:

$$Q_{\underbrace{l\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}'}_{\substack{\text{row} \\ l \in \{1,...,n\} \\ \bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\delta}}' \in \{0,2\}^n}}, \underbrace{i\boldsymbol{\delta}}_{\substack{\text{column} \\ i \in \{1,...,n\} \\ \boldsymbol{\delta} \in \{0,2\}^n}}} = \mathbb{1}\Big\{ \underbrace{(i = l)}_{\substack{\text{column} \\ \text{corresponds to } l}} \wedge \underbrace{(\bar{\boldsymbol{\delta}} \sqsupseteq_i \bar{\boldsymbol{\delta}}')}_{\substack{\bar{\boldsymbol{\delta}} \text{ influence-} \\ \text{dominates } \bar{\boldsymbol{\delta}}'}} \Big\} \times \Big(\underbrace{\mathbb{1}\{\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}\}}_{\substack{= 1 \text{ if column} \\ \text{corresponds to } \bar{\boldsymbol{\delta}}}} - \underbrace{\mathbb{1}\{\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}'\}}_{\substack{= 1 \text{ if column} \\ \text{corresponds to } \bar{\boldsymbol{\delta}}'}} \Big).$$

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 $\left[\begin{smallmatrix} \boldsymbol{v} \\ \boldsymbol{\theta} \end{smallmatrix}\right]$ satisfying*

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

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

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 integer weighted sum of rows of $\left[\begin{smallmatrix} P & X \\ 0 & Q \end{smallmatrix}\right]$ yields a vector of zeroes.

Condition 2. *There exist integer-valued vectors $\hat{\boldsymbol{p}} \in \mathbb{Z}^{\bar{m}}$, $\bar{\boldsymbol{p}} \in \mathbb{Z}^{m-\bar{m}}$, $\boldsymbol{q} \in \mathbb{Z}^{n2^n2^n}$ (with $\boldsymbol{p} \equiv \left[\begin{smallmatrix} \hat{\boldsymbol{p}} \\ \bar{\boldsymbol{p}} \end{smallmatrix}\right]$), satisfying:*

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

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

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 $\left[\begin{smallmatrix} P & X \\ \mathbf{0} & Q \end{smallmatrix}\right]$, which exactly cancel.

***

Our next Condition translates the matrix notation of Condition 2 into algebraic form:

Condition 3. *There exist vectors $\boldsymbol{p} \in \mathbb{N}^m$, $\boldsymbol{q} \in \mathbb{N}^{n2^n2^n}$, such that $\forall \boldsymbol{x} \in X$:*

$$\underbrace{\sum_{j:\{\boldsymbol{x}^j = \boldsymbol{x}\}}}_{\substack{\textit{appearances of } \boldsymbol{x} \\ \textit{on LHS}}} p_j = \underbrace{\sum_{j:\{\bar{\boldsymbol{x}}^j = \boldsymbol{x}\}}}_{\substack{\textit{appearances of } \boldsymbol{x} \\ \textit{on RHS}}} p_j,$$

and $\forall i \in \{1, \ldots, n\}, \boldsymbol{\delta} \in \Delta$,

$$\underbrace{\sum_{j:|x^j - z^j| = \boldsymbol{\delta}}}_{\substack{\textit{inequalities with} \\ \boldsymbol{\delta} \textit{ on LHS}}} p_j x_i^j \kappa_i \quad - \underbrace{\sum_{j:|x'^j - z'^j| = \boldsymbol{\delta}}}_{\substack{\textit{inequalities with} \\ \boldsymbol{\delta} \textit{ on RHS}}} p_j x_i'^j \kappa_i \quad + \underbrace{\sum_{\bar{\boldsymbol{\delta}}' : \boldsymbol{\delta} \sqsupseteq_i \bar{\boldsymbol{\delta}}'}}_{\substack{\textit{comparisons} \\ \textit{dominated by } \boldsymbol{\delta}}} q_{i\boldsymbol{\delta}\bar{\boldsymbol{\delta}}'} \quad - \underbrace{\sum_{\bar{\boldsymbol{\delta}} : \bar{\boldsymbol{\delta}} \sqsupseteq_i \boldsymbol{\delta}}}_{\substack{\textit{comparisons} \\ \textit{dominating } \boldsymbol{\delta}}} q_{i\bar{\boldsymbol{\delta}}\boldsymbol{\delta}} = 0$$

with $p_j > 0$ for some $j \leq \bar{m}$.

Proof that condition 3 is equivalent to condition 2. This proof is simply a rearrangement of the matrix condition of 2 into the algebraic conditions of 3.

First we can note that these two sets of assumptions are equivalent:

- $\hat{\boldsymbol{p}} \in \mathbb{Z}^{\bar{m}}$, $\bar{\boldsymbol{p}} \in \mathbb{Z}^{m-\bar{m}}$, $\boldsymbol{q} \in \mathbb{Z}^{n2^n2^n}$, $\hat{\boldsymbol{p}} > 0$, $\bar{\boldsymbol{p}} \geq 0$, $\boldsymbol{q} \geq 0$,
- $\boldsymbol{p} \in \mathbb{N}^m$, $\boldsymbol{q} \in \mathbb{N}^{n2^n2^n}$ with $p_j > 0$ for some $j \leq \bar{m}$.

Next, given $\boldsymbol{p}^T P = \mathbf{0}$ then, from the definition of P, we know that for each $\boldsymbol{x} \in \mathcal{X}$ we have:

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

Rearranging we get:

$$\sum_{j:\{\boldsymbol{x}^j = \boldsymbol{x}\}} p_j = \sum_{j:\{\bar{\boldsymbol{x}}^j = \boldsymbol{x}\}} p_j.$$

Next, given $\left[\, \boldsymbol{p}^T \; \boldsymbol{q}^T \,\right] \left[\begin{smallmatrix} X \\ Q \end{smallmatrix} \right] = \mathbf{0}^T$, then for each i and δ we have:

$$\underbrace{\sum_{j=1}^{m} p_j X_{j,i\boldsymbol{\delta}}}_{\text{elements of } X \text{ selected by } \boldsymbol{p}} \quad + \quad \underbrace{\sum_{l=1}^{n} \sum_{\bar{\boldsymbol{\delta}} \in \{0,2\}^n} \sum_{\bar{\boldsymbol{\delta}}' \in \{0,2\}^n} q_{l\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}'} Q_{l\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}',i\boldsymbol{\delta}}}_{\text{elements of Q selected by } \boldsymbol{q}} = 0$$

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

$$\sum_{j:|x^j - z^j| = \boldsymbol{\delta}} p_j x_i^j \kappa_i \quad - \sum_{j:|x'^j - z'^j| = \boldsymbol{\delta}} p_j x_i'^j \kappa_i \quad + \sum_{\bar{\boldsymbol{\delta}}':\boldsymbol{\delta} \sqsupseteq_i \bar{\boldsymbol{\delta}}'} q_{i\boldsymbol{\delta}\bar{\boldsymbol{\delta}}} \quad - \sum_{\bar{\boldsymbol{\delta}}:\bar{\boldsymbol{\delta}} \sqsupseteq_i \boldsymbol{\delta}} q_{i\bar{\boldsymbol{\delta}}\boldsymbol{\delta}} = 0$$

Finally, note that Condition 1 guarantees us only real-valued vectors $\boldsymbol{v}$ and $\boldsymbol{\theta}$, but our definition (1) imposes $\boldsymbol{\theta} \geq 0$. This is without loss of generality: for any utility function with real-valued $\boldsymbol{\theta}$ we can renormalize $v(.)$ and $\boldsymbol{\theta}$ to obtain a utility function with $\boldsymbol{\theta} \geq 0$ that assigns the same utility to every $(\boldsymbol{x}, \boldsymbol{z})$.

□

Proof of Theorem 2

We now prove that the matching condition in Theorem 2 is equivalent to Condition 2:

Condition 4. *There exists a cyclical selection $\boldsymbol{s} \in \mathbb{N}^m$ in which, (1) every attribute with a positive implicit preference ($\kappa_i = 1$) has an influence-negative matching, and (2) every attribute with a negative implicit preference ($\kappa_i = -1$) has an influence-positive matching.*

Proof that condition 4 implies condition 2. Given a cyclical selection $\boldsymbol{s}$ and a set of matching matrices $\{M_i\}_{i=1}^n$ we will construct vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ such that $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P & X\\ 0 & Q\end{smallmatrix}\right] = \boldsymbol{0}^T$.

Let:

$$\forall j \in \{1, \ldots, m\}, \qquad p_j = s_j$$
$$\forall i \in \{1, \ldots, n\},\ \boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0,2\}^n, \quad q_{i\boldsymbol{\delta}\boldsymbol{\delta}'} = M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'}\ .$$

By the definition of a cyclical selection we have $\hat{\boldsymbol{p}} > 0$ and $\bar{\boldsymbol{p}} \geq 0$, and by the definition of a matching we have $\boldsymbol{q} \geq 0$. For each element of the vector $\boldsymbol{p}^T P \in \mathbb{Z}^{2^n}$, which is indexed by $\boldsymbol{x}$, we can write:

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

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 $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}P\\ \boldsymbol{0}\end{smallmatrix}\right] = \boldsymbol{0}^T$.

An element of the vector $\left[\begin{smallmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{smallmatrix}\right]\left[\begin{smallmatrix}X\\ Q\end{smallmatrix}\right] \in \mathbb{Z}^{n2^n}$, indexed by $(i\boldsymbol{\delta})$, can be expressed as:

$$\sum_{j=1}^m p_j X_{j,i\boldsymbol{\delta}} + \qquad \sum_{l=1}^n \sum_{\bar{\boldsymbol{\delta}} \in \{0,2\}^n} \sum_{\bar{\boldsymbol{\delta}}' \in \{0,2\}^n} q_{l\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}'} Q_{l\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}',i\boldsymbol{\delta}}$$

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

$$\sum_{j:|x^j - z^j| = \boldsymbol{\delta}} p_j x_i^j \kappa_i \quad - \sum_{j:|x'^j - z'^j| = \boldsymbol{\delta}} p_j x_i'^j \kappa_i \quad + \sum_{\bar{\boldsymbol{\delta}}':\boldsymbol{\delta} \sqsupseteq_i \bar{\boldsymbol{\delta}}'} q_{i\boldsymbol{\delta}\bar{\boldsymbol{\delta}}'} \quad - \sum_{\bar{\boldsymbol{\delta}}:\bar{\boldsymbol{\delta}} \sqsupseteq_i \boldsymbol{\delta}} q_{i\bar{\boldsymbol{\delta}}\boldsymbol{\delta}} \tag{3}$$

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

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

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

$$\sum_{\bar{\boldsymbol{\delta}}':\boldsymbol{\delta} \sqsupseteq_i \bar{\boldsymbol{\delta}}'} M_{i,\boldsymbol{\delta},\bar{\boldsymbol{\delta}}'} - \sum_{\bar{\boldsymbol{\delta}}:\bar{\boldsymbol{\delta}} \sqsupseteq_i \boldsymbol{\delta}} M_{i,\bar{\boldsymbol{\delta}},\boldsymbol{\delta}} = \sum_{\bar{\boldsymbol{\delta}}' \in \{0,2\}^n} M_{i,\boldsymbol{\delta},\bar{\boldsymbol{\delta}}'} - \sum_{\bar{\boldsymbol{\delta}} \in \{0,2\}^n} M_{i,\bar{\boldsymbol{\delta}},\boldsymbol{\delta}} = \begin{cases} -c_{i,\boldsymbol{\delta}} & , \kappa_i = 1 \\ c_{i,\boldsymbol{\delta}} & , \kappa_i = -1 \end{cases}$$
$$= -\kappa_i c_{i,\boldsymbol{\delta}},$$

which uses Definition 8. The first equality follows from "matches obey dominance" and the second from "net flows are matched," by the premise that every attribute with $\kappa_i = 1$ has an influence-negative matching and every attribute with $\kappa_i = -1$ has an influence-positive matching. Substituting into (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 4. Given the vectors $\hat{\boldsymbol{p}}, \bar{\boldsymbol{p}}, \boldsymbol{q}$, we will construct a vector $\boldsymbol{s}$ and matrices $M_i, i = \{1, \ldots, n\}$ from and show that they satisfy Definitions 3 and 8:

$$\begin{aligned} \forall j \in \{1, \ldots, m\}, &\qquad s_j = p_j \\ \forall i \in \{1, \ldots, n\},\ \boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0,2\}^n, &\quad M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'} = q_{i\boldsymbol{\delta}\boldsymbol{\delta}'} \mathbb{1}\{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'\} \end{aligned}$$

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 $M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'} \geq 0$ 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 $\boldsymbol{p}^T P = \mathbf{0}^T$ (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 M_i satisfies the conditions of Definition 8. Observe that:

1. Matches obey dominance: $\forall \boldsymbol{\delta}, \boldsymbol{\delta}' \in \{0,2\}^n,\ (M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'} > 0) \implies (\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}')$. This immediately follows because we constructed M_i from $\boldsymbol{q}$ as $M_{i,\boldsymbol{\delta},\boldsymbol{\delta}'} = q_{i\boldsymbol{\delta}\boldsymbol{\delta}'} \mathbb{1}\{\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}'\}$.
2. Net flows are matched, i.e. for every $\boldsymbol{\delta} \in \{0,2\}^n$ and $i \in \{1, \ldots, n\}$ with $\kappa_i = 1$:

$$\begin{aligned} \sum_{\bar{\boldsymbol{\delta}}' \in \{0,2\}^n} M_{i,\boldsymbol{\delta},\bar{\boldsymbol{\delta}}'} - \sum_{\bar{\boldsymbol{\delta}} \in \{0,2\}^n} M_{i,\bar{\boldsymbol{\delta}},\boldsymbol{\delta}} &= \sum_{\bar{\boldsymbol{\delta}}' : \boldsymbol{\delta} \sqsupseteq_i \bar{\boldsymbol{\delta}}'} q_{i,\boldsymbol{\delta},\bar{\boldsymbol{\delta}}'} - \sum_{\bar{\boldsymbol{\delta}} : \bar{\boldsymbol{\delta}} \sqsupseteq_i \boldsymbol{\delta}} q_{i,\bar{\boldsymbol{\delta}},\boldsymbol{\delta}} && \text{(by construction of } M\text{)} \\ &= (\boldsymbol{q}^T Q)_{i\boldsymbol{\delta}} && \text{(by definition of } Q\text{)} \\ &= -(\boldsymbol{p}^T X)_{i\boldsymbol{\delta}} && \text{(by condition 2)} \\ &= -\sum_{j : |\boldsymbol{x}^j - \boldsymbol{z}^j| = \boldsymbol{\delta}} p_j x_i^j + \sum_{j : |\boldsymbol{x}'^j - \boldsymbol{z}'^j| = \boldsymbol{\delta}} p_j x_i'^j && \text{(by definition of } X \text{ and } \kappa_i = 1\text{)} \\ &= -\sum_{j : |\boldsymbol{x}^j - \boldsymbol{z}^j| = \boldsymbol{\delta}} s_j x_i^j + \sum_{j : |\boldsymbol{x}'^j - \boldsymbol{z}'^j| = \boldsymbol{\delta}} s_j x_i'^j && \text{(by construction of } \boldsymbol{s}\text{)} \\ &= -c_{i,\boldsymbol{\delta}} && \text{(by definition of } c_{i,\boldsymbol{\delta}}\text{)} \end{aligned}$$

So i has an influence-negative matching when $\kappa_i = 1$. The same argument will show that when $\kappa_i = -1$, then i has an influence-positive matching.

□

References

Ahumada, B., Y. Chen, N. Gupta, K. Hyde, M. Lepper, W. Mathews, N. Silveus, L. Vesterlund, T. Weidman, A. Wilson, K. P. Winichakul, and L. Zhou (2022). Well excuse me! replicating and connecting excuse-seeking behaviors. *Economics: Faculty Publications, Smith College*.

Alesina, A., M. Carlana, E. L. Ferrara, and P. Pinotti (2018). Revealing Stereotypes: Evidence from Immigrants in Schools. *American Economic Review, forthcoming*.

Allen, R. and J. Rehbeck (2023). Revealed stochastic choice with attributes. *Economic Theory 75*(1), 91–112.

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. *Management Science, forthcoming*.

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 (2023). Inaccurate Statistical Discrimination: An Identification Problem. *Review of Economics and Statistics, forthcoming*.

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. `https://kcborder.caltech.edu/Notes/Alternative.pdf` *accessed 2022-04-24*.

Bursztyn, L., G. Egorov, I. Haaland, A. Rao, and C. Roth (2023). Justifying dissent. *The Quarterly Journal of Economics 138*(3), 1403–1451.

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.

Corno, L., E. L. Ferrara, and J. Burns (2022). Interaction, stereotypes, and performance: Evidence from south africa. *American Economic Review 112*(12), 3848–3875.

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. (2013a). 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.

DeSante, C. D. (2013b). Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor: Dataset. *Harvard Dataverse,* `https://doi.org/10.7910/DVN/AZTWDW`.

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. (2016a). Excusing Selfishness in Charitable Giving: The Role of Risk. *The Review of Economic Studies 83*(2), 587–628.

Exley, C. L. (2016b). Excusing Selfishness in Charitable Giving: The Role of Risk: Dataset. *The Review of Economic Studies,* `https://doi.org/10.1093/restud/rdv051`.

Fishburn, P. C. (1970). *Utility theory for decision making.* Krieger NY.

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.

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

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kahneman, D. and S. Frederick (2005). A Model of Heuristic Judgment. *The Cambridge handbook of thinking and reasoning*, 267–294.

Kessler, J. B., C. Low, and X. Shan (2023). Lowering the playing field: Discrimination through sequential spillover effects. *Review of Economics and Statistics, forthcoming*.

Kőszegi, B. and A. Szeidl (2012). A Model of Focusing in Economic Choice. *The Quarterly Journal of Economics 128*(1), 53–104.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy 74*(2), pp. 132–157.

Loewenstein, G. F., E. U. Weber, C. K. Hsee, and N. Welch (2001). Risk as feelings. *Psychological Bulletin 127*(2), 267–286.

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.

Wakker, P. P. (1989). *Additive representations of preferences: A new foundation of decision analysis*, Volume 4. Springer Science & Business Media.

Wernerfelt, B. (1995). A rational reconstruction of the compromise effect: Using market data to infer utilities. *Journal of Consumer Research 21*(4), 627–633.

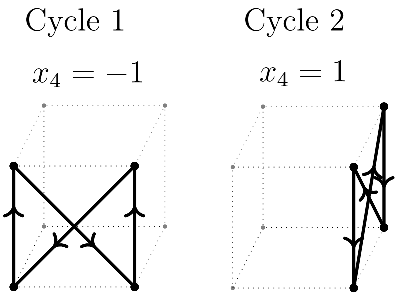
Woodford, M. (2012). Inattentive Valuation and Reference-Dependent Choice. *mimeo*.

A Web Appendix to "Implicit Preferences"

For Online Publication Only

A.1 Example dataset that falsifies Dilution but not Cancellation

Consider a decision maker choosing between bundles with $n = 4$ attributes. In 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 $x_4 = 1$.



Observed comparisons

$$\begin{aligned}\delta_V &= [0,2,0,0]^T \quad \text{(vertical)}\\ \delta_{FD} &= [2,2,0,0]^T \quad \text{(front diagonal)}\\ \delta_{SD} &= [0,2,2,0]^T \quad \text{(side diagonal)}\end{aligned}$$

Denote by s^1 the weight put on cycle 1 in cyclical selection $\boldsymbol{s}$, and s^2 the weight on cycle 2. At least one must be strictly positive. The condition in Proposition 1 tells us that a dataset is consistent with a separable comparative utility function provided there is no $\boldsymbol{s}$ such that every entry in the score vector $\boldsymbol{c}$ is nonzero. Observe that for attribute 2 and $\boldsymbol{\delta}_{FD}$ the score is $4s^1$, while for attribute 2 and $\boldsymbol{\delta}_{SD}$ the score is $-4s^2$, therefore at least one of these entries must be nonzero in any cyclical selection: there exists a separable comparative utility function consistent with the dataset.

However that separable comparative utility function does not satisfy Dilution (Assumption 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 $\boldsymbol{\kappa}$ that can rationalize the dataset.

A.2 "Identification Cookbook": Derivation of Canonical Examples

We now formally define each of the examples from Section 2, and derive their implications using our representation results. All definitions are stated in terms of strict inequalities, but it is sufficient if at least one inequality in each cyclical selection is strict. We always assume Dilution (Assumption 1) holds, and introduce Dominance-k (Assumption 2) when we get to evaluation.

We recommend reading the definitions alongside the diagrams in Section 2.

Choice Examples.

Definition 13 (Right triangle). *A right triangle is a choice cycle over three distinct bundles, ordered $\boldsymbol{x}^1 \succ \boldsymbol{x}^2 \succ \boldsymbol{x}^3 \succ \boldsymbol{x}^1$, where $|\boldsymbol{x}^1 - \boldsymbol{x}^2| + |\boldsymbol{x}^2 - \boldsymbol{x}^3| = |\boldsymbol{x}^1 - \boldsymbol{x}^3|$.*

Corollary 3. *A right triangle reveals at least one non-zero implicit preference, favoring $\boldsymbol{x}^3$'s realization of an attribute that is non-shared with $\boldsymbol{x}^1$:*

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

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.

Definition 14 (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 comparison vectors $|\boldsymbol{x}^1 - \boldsymbol{x}^2| = |\boldsymbol{x}^3 - \boldsymbol{x}^4|$, 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 4. *A figure 8 reveals at least one non-zero implicit preference, favoring $\boldsymbol{x}^4$'s realization of an attribute that is non-shared with $\boldsymbol{x}^3$:*

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

When $\boldsymbol{x}^3$ and $\boldsymbol{x}^4$ differ on a single attribute, we learn the sign of its implicit preference. The figure 8 is the shortest choice cycle that can isolate a single implicit preference.

Definition 15 (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 $\bar{\boldsymbol{x}}^1 \succ \bar{\boldsymbol{x}}^2 \succ \bar{\boldsymbol{x}}^3 \succ \bar{\boldsymbol{x}}^1$, satisfing 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 $(\bar{\boldsymbol{x}}^2, \bar{\boldsymbol{x}}^3)$ $(\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3))$.*

Corollary 5. *A pair of parallel right triangles reveals at least one non-zero implicit preference, favoring $\boldsymbol{x}^3$'s realization of an attribute that is non-shared with $\boldsymbol{x}^2$:*

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

Parallel right triangles refine the identification of single triangles 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 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).

Evaluation Examples. With evaluation data we cannot ignore implicit preferences on shared attributes.As a result, in general we identify disjunctions over implicit preferences on *every* attribute. Additionally, Dilution does not restrict how influence changes when an attribute goes from shared to non-shared, so we sometimes draw indeterminate conclusions about some implicit preferences. Assumption 2 can resolve these indeterminacies.

The cyclical selections we study consist of one or more "scissor" inequalities, which have the same target on the left- and right-hand sides, taking the form $u(\boldsymbol{x}, \boldsymbol{z}) > u(\boldsymbol{x}, \boldsymbol{z}')$.[46]

Definition 16 (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 6. *A convex scissor reveals at least one non-zero implicit preference:*

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

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

[46] Because of how we construct evaluation datasets (see section 1), this inequality may not literally appear in the dataset. E.g., 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. We can construct an equivalent cyclical selection by including every inequality that lies in between. This is equivalent to the single inequality, because every intermediate $u(\boldsymbol{x}'', \boldsymbol{z}'')$ will appear on the RHS of one inequality and the LHS of the next, so their contributions to the score for the corresponding $\boldsymbol{\delta}$ will always equal zero.

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, while attributes that are shared in both comparisons become *less* dilute. Attributes that are shared in $|\boldsymbol{x}-\boldsymbol{z}^1|$ but non-shared in $|\boldsymbol{x}-\boldsymbol{z}^2|$ change influence but Assumption 1 (Dilution) does not tell us in which direction.

Assumption 2 (Dominance-k) resolves the ambiguity. While the assumption allows for either i or k to change status, we assume for simplicity that attribute k is either always shared (i.e., $k \in \{i : x_i = z_i^2\}$), or always non-shared (i.e., $k \in \{i : x_i \neq z_i^1\}$), in all comparisons.

Corollary 7 (Convex scissor with Dominance-k)**.** *Suppose Assumption 2 holds. Let* $\Theta = 1$ *when* k *is always shared* ($x_k = z_k^1 = z_k^2$)*, and* $\Theta = -1$ *when* k *is always non-shared* ($x_k \neq z_k^1$ *and* $x_k \neq z_k^2$)*. A convex scissor reveals:*

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

The examples in section 2 assumed attribute k is shared ($\Theta = 1$).

Combining two scissors can refine our identification of implicit preferences:

Definition 17 (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.,* $|\boldsymbol{x} - \boldsymbol{z}^2| = |\bar{\boldsymbol{x}} - \bar{\boldsymbol{z}}^2|$*).*[47]

Corollary 8. *A pair of parallel convex scissors reveals at least one non-zero implicit preference. There are many possible cases, summarized in the following disjunction:*

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

Parallel convex scissors refine the implications of their constituent scissors when there are attributes with $x_i\Upsilon = -\bar{x}_i\bar{\Upsilon}$, because the terms associated with those attributes are eliminated. Note that this does not rely on Assumption 2. Intuitively, the observed behavior cannot be exclusively driven by any combination of implicit preferences on eliminated attributes, because evaluation moved in contradictory directions in the two scissors. We achieve

[47] 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.

unique identification if all but one attribute is eliminated. Assumption 2 can further refine identification when some implicit preferences are inconclusive:

Corollary 9 (Parallel convex scissors with Dominance-k)**.** *Suppose Assumption 2 holds. Let* $\Theta = 1$ *when k is always 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 is always non-shared* ($x_k \neq z_k^1 = z_k^2$) *and* ($\bar{x}_k \neq \bar{z}_k^1 = \bar{z}_k^2$)*. The parallel convex scissors reveal:*

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

A.2.1 Proofs

Our derivations use a couple of tricks. First we show that we can represent each example with just three "grouped" attributes. Second we express each cycle and each relevant restriction from Assumptions 1 and 2 as simplified X and Q matrices (X and Q are defined in the proof of Theorem 1). All possible matchings can be expressed as a linear combination of rows of these matrices. From this we easily deduce which realizations of $\boldsymbol{\kappa}$ are ruled out.

Reduction to three attributes. All of our examples can be analyzed by partitioning the attribute space into three disjoint and collectively exhaustive "groups," A, B, C. All attributes within a group are perfectly correlated, so we can represent them using three grouped attributes, $\boldsymbol{x} = (x_A, x_B, x_C)$.[48] Since attributes are perfectly correlated within groups, they will have identical differences in a given comparison (e.g. we have $|x_i - z_i| = |x_j - z_j|, \forall i, j \in A$). All influence-dominance relationships will be shared within a group (e.g. $(\boldsymbol{\delta} \sqsupseteq_i \boldsymbol{\delta}') \Leftrightarrow (\boldsymbol{\delta} \sqsupseteq_j \boldsymbol{\delta}'), \forall i, j \in A$). 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: $(x_A\kappa_A = 1) \implies \left(\bigvee_{i\in A} x_i\kappa_i = 1\right)$.

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 $\left[\, \boldsymbol{p}^T \; \boldsymbol{q}^T \,\right] \left[\begin{smallmatrix} P & X \\ 0 & Q \end{smallmatrix}\right] = \mathbf{0}$. 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 at least one term must be nonzero. We use a number of tricks to simplify the process.

[48] So, $A \cup B \cup C = 1, \ldots, n$; 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 $x_A = 1 \Leftrightarrow (x_1, x_2, x_3) = (1, -1, 1)$.

$$
\begin{array}{rc|ccccccccc|}
 & & A,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & A,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & A,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & B,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & B,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & B,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & C,\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & C,\left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & C,\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \\
 & \text{Right triangle 1} & -2\kappa_A x_A^3 & 0 & 2\kappa_A x_A^3 & 0 & -2\kappa_B x_B^3 & 2\kappa_B x_B^3 & 0 & 0 & 0 \\
 & \text{Right triangle 2} & -2\kappa_A \bar{x}_A^3 & 0 & 2\kappa_A \bar{x}_A^3 & 0 & -2\kappa_B \bar{x}_B^3 & 2\kappa_B \bar{x}_B^3 & 0 & 0 & 0 \\
X^* = & \text{Figure 8} & -4\kappa_A x_A^4 & 0 & 4\kappa_A x_A^4 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & \text{Convex scissor 1} & -\kappa_A x_A \Upsilon & 0 & \kappa_A x_A \Upsilon & -\kappa_B x_B \Upsilon & 0 & \kappa_B x_B \Upsilon & -\kappa_C x_C \Upsilon & 0 & \kappa_C x_C \Upsilon \\
 & \text{Convex scissor 2} & -\kappa_A \bar{x}_A \bar{\Upsilon} & 0 & \kappa_A \bar{x}_A \bar{\Upsilon} & -\kappa_B \bar{x}_B \bar{\Upsilon} & 0 & \kappa_B \bar{x}_B \bar{\Upsilon} & -\kappa_C \bar{x}_C \bar{\Upsilon} & 0 & \kappa_C \bar{x}_C \bar{\Upsilon}
\end{array}
$$

$$
\begin{array}{rc|ccccccccc|}
 & \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \sqsupseteq_A \left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
Q^* = & \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] \sqsupseteq_B \left[\begin{smallmatrix}0\\2\\0\end{smallmatrix}\right] & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
 & \left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right] \sqsupseteq_C \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right] & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
 & \text{Dominance-}k & 0 & 0 & 0 & \Theta & 0 & -\Theta & 0 & 0 & 0
\end{array}
$$

Notes: (1) Columns are labeled by attribute group ($i \in \{A, B, C\}$), and $|\boldsymbol{x} - \boldsymbol{z}| \in \{0, 2\}^3$. (2) Row elements of X^* correspond to κ_i multiplied by the score. (3) Entries in Q^* are derived from Assumptions 1 and 2. We only include rows that restrict at least one row of X^*. (4) For scissors, $\Upsilon = sgn(y^2 - y^1) \in \{-1, 1\}$ equals the sign of the evaluation change. (5) $\Theta \in \{-1, 0, 1\}$ captures the sign of the Dominance-k assumption (Assumption 2). $\Theta = 0$ if the assumption does not apply, $\Theta = 1$ if influence is higher for shared attributes (attribute k is shared), $\Theta = -1$ if influence is higher for non-shared (k is non-shared).

Figure A1: Matrix representation of corollaries

First, we can ignore P, since in any solution, $\boldsymbol{p}$ is a cyclical selection and P's rows always sum to zero in a cyclical selection. Thus we focus on X and Q.

Second, many comparisons $\boldsymbol{\delta} \in \{0, 2\}^n$ are never observed (i.e., are not in the dataset), so appear in X as columns of zeros. We drop those columns. Similarly, Q will have many rows that do not restrict any nonzero column in X. We eliminate those as well. We therefore write X and Q with one column per observed realization of $|\boldsymbol{x} - \boldsymbol{z}|$.

Third, for a dataset consisting of a single 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 left- and right-hand sides of the selected inequalities. That means all p_j terms must be equal for inequalities in the cycle. When the dataset consists of multiple cycles we collapse each down to a single row.

We name the compressed X and Q matrices X^* and Q^* and show them in Figure A1.

Calculating scores: Entries in X^* correspond to *scores*, which are the net of "wins" and "losses" for each i and $\boldsymbol{\delta}$. We count a "win" for $i, \boldsymbol{\delta}$ for each inequality j in which $|\boldsymbol{x}^j - \boldsymbol{z}^j| = \boldsymbol{\delta}$ and $x_i^j = 1$, and for each inequality j in which $|\boldsymbol{x}' - \boldsymbol{z}'| = \boldsymbol{\delta}$ and $x_i'^j = -1$ (i.e., a positive value of x_i appears on the LHS of an inequality, or a negative value on the RHS). "Losses" correspond to the opposite case. So, the inequality $u\left(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}-1\\-1\end{smallmatrix}\right]\right) > u\left(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]\right)$ contributes a win for $1, \left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$ and for $2, \left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$, and contributes a loss for $1, \left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$ and for $2, \left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$.

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 $|\boldsymbol{x}^1 - \boldsymbol{x}^2| + |\boldsymbol{x}^2 - \boldsymbol{x}^3| = |\boldsymbol{x}^1 - \boldsymbol{x}^3|$, A, B, C are disjoint and collectively exhaustive.

There are three inequalities in the dataset: $u(\boldsymbol{x}^1, \boldsymbol{x}^2) > u(\boldsymbol{x}^2, \boldsymbol{x}^1)$, $u(\boldsymbol{x}^2, \boldsymbol{x}^3) > u(\boldsymbol{x}^3, \boldsymbol{x}^2)$, and $u(\boldsymbol{x}^3, \boldsymbol{x}^1) > u(\boldsymbol{x}^1, \boldsymbol{x}^3)$. The comparisons are: $|\boldsymbol{x}^1 - \boldsymbol{x}^2| = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$, $|\boldsymbol{x}^2 - \boldsymbol{x}^3| = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$, and $|\boldsymbol{x}^1 - \boldsymbol{x}^3| = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$. To construct the X^* matrix we need to compute the score for each $i, |\boldsymbol{x} - \boldsymbol{z}|$.

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 $x_C^1 = x_C^2 = x_C^3$. 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 $2\kappa_A x_A^1 = -2\kappa_A x_A^3$ (since $x_A^1 = -x_A^3$). All other attributes are shared so have zero score.

Inequality 2 yields two wins for B, $\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$ if $x_B^2 = 1$ and two losses otherwise, so that column's entry is $2\kappa_B x_B^2 = -2\kappa_B x_B^3$ (since $x_B^2 = -x_B^3$). All other attributes are shared.

Inequality 3 gives us two wins for A, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ if $x_A^3 = 1$ and two losses if $x_A^3 = -1$, so that column's entry is $2\kappa_A x_A^3$. Inequality 3 gives us two wins for B, $\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$ if $x_B^3 = 1$ and two losses if $x_B^3 = -1$, so so that column's entry is $2\kappa_B x_B^3$. All other attributes are shared.

Adding rows, we obtain "Right triangle 1" in Figure A1. Our dataset has just one cycle so we can set $\boldsymbol{p} = 1$ without loss of generality. Dropping columns that equal zero, we obtain:

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

where q_1, q_2 are the coefficients on the first and second rows of Q^* respectively. There exist $q_1, q_2 \geq 0$ such that this vector equals $\mathbf{0}$ if and only if $(\kappa_A x_A^3 \leq 0) \wedge (\kappa_B x_B^3 \leq 0)$. This is false 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. A, B, C are disjoint and collectively exhaustive.

The two comparisons are $|\boldsymbol{x}^1 - \boldsymbol{x}^2| = |\boldsymbol{x}^3 - \boldsymbol{x}^4| = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$, 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. Once again we can set $\boldsymbol{p} = 1$ and eliminate columns of zeros without loss of generality, obtaining:

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

where q_1 is the coefficient on the first row of Q^*. This vector is nonzero if and only if:

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

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

Parallel right triangles Let $A = \{i : x_i^1 \neq x_i^2\} = \{i : \bar{x}_i^2 \neq \bar{x}_i^3\}$, $B = \{i : x_i^2 \neq x_i^3\} = \{i : \bar{x}_i^1 \neq \bar{x}_i^2\}$, $C = \{i : x_i^1 = x_i^3\} = \{i : \bar{x}_i^1 = \bar{x}_i^3\}$. In words, A is the set of non-shared attributes in $|\boldsymbol{x}^1 - \boldsymbol{x}^2|$ and $|\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3|$, B is the set of non-shared attributes in $|\boldsymbol{x}^2 - \boldsymbol{x}^3|$ and $|\bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2|$, while C attributes are always shared.[49] A, B, C are disjoint and collectively exhaustive.

We populate Right Triangle 2's row in X^* exactly as we did for Right Triangle 1. 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):

$$\begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{q}^T \end{bmatrix} \left[\begin{smallmatrix} X^* \\ Q^* \end{smallmatrix}\right] = \begin{bmatrix} -W_A & W_A & -W_B & W_B \end{bmatrix} \tag{4}$$

$$W_A = 2\kappa_A(p_1 x_A^3 + p_2 \bar{x}_A^3) + q_1 = 2\kappa_A(p_1 - p_2)x_A^3 + q_1 \tag{5}$$

$$W_2 = 2\kappa_B(p_1 x_B^3 + p_2 \bar{x}_B^3) + q_2 = 2\kappa_B(p_1 + p_2)x_B^3 + q_2, \tag{6}$$

where q_1, q_2 are the coefficients on the first and second rows of Q^*, respectively. The second steps in (5) and (6) use $x_A^3 = -\bar{x}_A^3$, and $x_B^3 = \bar{x}_B^3$.[50] (4) is non-zero if and only if:

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

This condition must hold for all $\boldsymbol{p}$s. Observe that it is sufficient to check the case where $p_1 = p_2$ (where the cyclical selection contains an equal number of each cycle). Then, the

[49] While C attributes are always shared within a triangle, they could take different realizations between triangles; i.e. we could have $\boldsymbol{x}_C \neq \bar{\boldsymbol{x}}_C$.

[50] Derivation: Condition 1 in the definition of the parallel right triangle ($\boldsymbol{x}^2 - \boldsymbol{x}^3 = \bar{\boldsymbol{x}}^1 - \bar{\boldsymbol{x}}^2$) allows us to pin down the values of the non-shared attributes in 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 $\boldsymbol{x}_B^3 = -1$). Condition 2 in the definition ($\boldsymbol{x}^1 - \boldsymbol{x}^2 = -(\bar{\boldsymbol{x}}^2 - \bar{\boldsymbol{x}}^3)$) allows us to pin down the values of the non-shared attributes in 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 $x_B^3 = \bar{x}_B^3$.

condition reduces to $\kappa_B x_B^3 = 1$. By the definition of set B we conclude that a pair of parallel right triangles reveals:

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

Convex scissor without and with Dominance-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 non-shared attributes in the first comparison, B is the set of attributes that are non-shared in the second comparison but shared in the first, $A \cup B$ the full set that vary in the second comparison, and C the set that are always shared. A, B, C are disjoint and collectively exhaustive.

A scissor can be written as a single inequality (see footnote 46). We construct its row in X^* by computing scores as usual. 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 $|\boldsymbol{x} - \boldsymbol{z}^2| = \left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right]$, giving us a win in column i, $\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right]$ if $x_i = 1$ and a loss otherwise. The right-hand side corresponds to $|\boldsymbol{x} - \boldsymbol{z}^1| = \left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right]$, giving us a loss in column i, $\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right]$ if $x_i = 1$ and a win otherwise. If $y^2 < y^1$ the LHS and RHS are switched. Defining $\Upsilon = sgn(y^2 - y^1)$, we enter $\kappa_i x_i \Upsilon$ in the columns for comparison $\left[\begin{smallmatrix}2\\2\\0\end{smallmatrix}\right]$, and $-\kappa_i x_i \Upsilon$ in the columns for comparison $\left[\begin{smallmatrix}2\\0\\0\end{smallmatrix}\right]$. Setting $\boldsymbol{p} = 1$ (wlog), and ignoring zeros, we obtain:

$$\begin{aligned} \begin{bmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{bmatrix} \left[\begin{smallmatrix}X^*\\Q^*\end{smallmatrix}\right] &= \begin{bmatrix} -W_A & W_A & -W_B & W_B & -W_C & W_C\end{bmatrix} \\ W_A &= \kappa_A x_A \Upsilon + q_1 \\ W_B &= \kappa_B x_B \Upsilon - \Theta q_4 \\ W_C &= \kappa_C x_C \Upsilon - q_3. \end{aligned} \tag{7}$$

q_1, q_3, q_4 are the coefficients on the first, third, and fourth rows of Q^*. Θ encodes the Dominance-k assumption (Assumption 2): $\Theta = 0$ if there is no attribute k for which the assumption holds, $\Theta = 1$ 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 $\Theta = 0$, (7) is non-zero if and only if:

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

while when $\Theta \neq 0$, (7) is non-zero 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. $\square$

Parallel convex scissors without and with Dominance-k. Conditions 1 and 2 in the definition 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: $A = \{i : x_i \neq z_i^1\} = \{i : \bar{x}_i \neq \bar{z}_i^1\}$, $B = \{i : z_i^1 \neq z_i^2\} = \{i : \bar{z}_i^1 \neq \bar{z}_i^2\}$, and $C = \{i : x_i = z_i^2\} = \{i : \bar{x}_i = \bar{z}_i^2\}$.

So, A is the set of non-shared attributes in each scissor's the first comparison, B is the set of attributes that are non-shared in their second comparisons but shared in the first, $A \cup B$ the full set that vary in the second comparisons, and C the set that are always shared. A, B, C are disjoint and collectively exhaustive. The values of $\boldsymbol{x}, \bar{\boldsymbol{x}}, sgn(y^2 - y^1)$ and $sgn(\bar{y}^2 - \bar{y}^1)$ are unrestricted, so there are many possible combinations.

A cyclical selection consists of $p_1 \geq 0$ copies of the first scissor and $p_2 \geq 0$ copies of the second, giving us (ignoring zero elements):

$$\begin{bmatrix}\boldsymbol{p}^T & \boldsymbol{q}^T\end{bmatrix} \left[\begin{smallmatrix} X^* \\ Q^* \end{smallmatrix}\right] = \begin{bmatrix}-W_A & W_A & -W_B & W_B & -W_C & W_C\end{bmatrix} \tag{8}$$
$$W_A = \kappa_A(p_1 x_A \Upsilon + p_2 \bar{x}_A \bar{\Upsilon}) + q_1$$
$$W_2 = \kappa_B(p_1 x_B \Upsilon + p_2 \bar{x}_B \bar{\Upsilon}) - \Theta q_4$$
$$W_3 = \kappa_C(p_1 x_C \Upsilon + p_2 \bar{x}_C \bar{\Upsilon}) - q_3,$$

q_1, q_3, q_4 are the coefficients on the first, third, and fourth rows of Q^*. Θ encodes the Dominance-k assumption (Assumption 2) as before. $\Upsilon = sgn(y^2 - y^1)$ and $\bar{\Upsilon} = sgn(\bar{y}^2 - \bar{y}^1)$ capture the directions in which each evaluation changes when the comparator changes.

By a similar argument to the parallel right triangles, it is sufficient to check the case where $p_1 = p_2$. 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 $\Theta = 0$, (8) is non-zero if and only if:

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

while when $\Theta \neq 0$, (8) is non-zero if and only if:

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

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

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. Some parallel scissors eliminate attribute group B (where a single scissor is indeterminate), enabling us to draw precise conclusions without invoking Assumption 2.

A.3 Proofs for Section 3 (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 that $S^{|\boldsymbol{x}-\boldsymbol{z}|} = \{i : |x_i - z_i| = 0\}$.

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

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

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

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

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

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 $Var\left[\begin{smallmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{smallmatrix}\right] = \left[\begin{smallmatrix}\Sigma_a & \Sigma_{a,b} \\ \Sigma_{a,b}^T & \Sigma_b\end{smallmatrix}\right]$ we can write the conditional expectation: $E[\boldsymbol{a}|\boldsymbol{b}] = E[\boldsymbol{a}] + \Sigma_{a,b}\Sigma_b^{-1}(\boldsymbol{b} - E[\boldsymbol{b}])$. In our case this implies:

$$E[\boldsymbol{w}|\boldsymbol{y}] = \Sigma_{w,y}\Sigma_y^{-1}\boldsymbol{y} \tag{9}$$

with components as follows:

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

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

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

Thus, given (9), we obtain:

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

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 also applies to the cases of all shared attributes ($\boldsymbol{x} = \boldsymbol{z}$) and all non-shared attributes ($\boldsymbol{x} = -\boldsymbol{z}$). We cannot use equation (9) because X does not have full rank so Σ_y is not invertible. If all attributes are shared 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, $E[w_i|\boldsymbol{y}] = x_i\dfrac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2}\bar{y}^x = x_i\dfrac{\sigma_i^2}{\sum_{j\in S}\sigma_j^2}\dfrac{\bar{y}^x+\bar{y}^z}{2}$. If all attributes are non-shared 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. □

Proof Strategy for Propositions 2–6. To prove Propositions 2–5 we will show that the utility function defined in each foundation can be expressed as an Implicit Preferences utility function. In each case, θ_i takes one of two functional forms depending on whether i is shared or non-shared:

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

To verify Dilution we show that $\theta_i(|\boldsymbol{x}-\boldsymbol{z}|)$ is weakly increasing as the set of attributes that have the same status as 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 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 6, for each foundation we write out the change in θ_i when i's status becomes the same as k's, and show it is weakly positive.

Proof of Proposition 2. 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^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 $\theta_i^N(|\boldsymbol{x} - \boldsymbol{z}'|) = |\lambda_i| \geq \theta_i^N(|\boldsymbol{x} - \boldsymbol{z}|)$. □

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

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):

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

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

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. This concludes the proof. □

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:

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

Hence $y^x(\boldsymbol{x}, \boldsymbol{z}) = u^{SE}(\boldsymbol{x}, \boldsymbol{z})$ and $y^z(\boldsymbol{z}, \boldsymbol{x}) = u^{SE}(\boldsymbol{z}, \boldsymbol{x})$ as stated in the proposition.
Next we show that player 2's strategy is optimal, given player 1's. Taking first order conditions of U^2, and using Lemma 2, we obtain the desired result:

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

Proof of Proposition 4. Using the fact that $\lambda_i = sgn(\lambda_i)|\lambda_i|$ we can express $u^{SE}(|\boldsymbol{x} - \boldsymbol{z}|)$ in a form that satisfies (1):

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

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

It is easy to see that θ^S and θ^N are weakly increasing as we add additional shared and non-shared attributes respectively. □

Proof of Proposition 5. The result in brief: when attributes 1 and 2 have different status, Agent 2 can perfectly infer Agent 1's private information (λ_1, λ_2) from their reports

$(\hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z}))$. When they have the same status, Agent 2 can only infer a weighted "average" λ. All other attributes have no effect because agent 1 has no private information about them. This means dilution will be satisfied: the influence of attribute i (weakly) increases when the set of attributes with the same status as i grows.[51]

Agent 1's reported value for $\boldsymbol{x}$ is $\hat{f}(\boldsymbol{x}) = E[f(\boldsymbol{x})|\boldsymbol{\lambda}]$. Given agent 1's prior ($E[\boldsymbol{\pi}] = \mathbf{1}$), we have

$$\hat{f}(\boldsymbol{x}) = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i E[\pi_i] = g(\boldsymbol{x}) + \sum_{i=1}^{n} x_i \lambda_i = g(\boldsymbol{x}) + x_1\lambda_1 + x_2\lambda_2.$$

where the final step follows from $\sigma_i^2 = 0, \forall i > 2$ (implying $\lambda_i = 0, \forall i > 2$).

Define the residual value by subtracting common-knowledge components: $\bar{\hat{f}}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) = x_1\lambda_1 + x_2\lambda_2$. We can then use Lemma 2 to express Agent 2's posterior on λ_i in a simple form (we use $-i$ to denote the other attribute $j \in \{1,2\} \setminus i$):

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

$$= \begin{cases} x_i \frac{\sigma_i^2}{\sigma_i^2+\sigma_{-i}^2}(x_i\lambda_i + x_{-i}\lambda_{-i}) & , \delta_i = \delta_{-i} \\ \lambda_i & , \delta_i \neq \delta_{-i} \end{cases}$$

Substituting this into $u^{IA}(\boldsymbol{x}, \boldsymbol{z}) = E[f(\boldsymbol{x})|\boldsymbol{\pi}, \hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})]$, and rearranging, we obtain:

$$E[f(\boldsymbol{x})|\boldsymbol{\pi}, \hat{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{z})] = \begin{cases} g(\boldsymbol{x}) + x_1\lambda_1 \frac{\pi_1\sigma_1^2+\pi_2\sigma_2^2}{\sigma_1^2+\sigma_2^2} + x_2\lambda_2 \frac{\pi_1\sigma_1^2+\pi_2\sigma_2^2}{\sigma_1^2+\sigma_2^2} & , \delta_i = \delta_{-i} \\ g(\boldsymbol{x}) + x_1\lambda_1\pi_1 + x_2\lambda_2\pi_2 & , \delta_i \neq \delta_{-i} \end{cases}$$

We can see that if attributes 1 and 2 have the same status (both shared or both non-shared) then both are weighted by the average π (i.e. $\frac{\pi_1\sigma_1^2+\pi_2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$). If they do not have the same status

[51] When all σ_i^2 can be nonzero, it is possible to show that Agent 2's utility can be written as:

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

The weight on $x_i\lambda_i$ is a weighted *average* π_j across all j with the same status. Thus, Dilution is not guaranteed to hold, because this average can either increase or decrease when the set of same-status attributes grows.

there is full revelation and each is weighted by its own π. We can therefore write:

$$u^{IA}(\boldsymbol{x},\boldsymbol{z}) = \underbrace{g(\boldsymbol{x}) + \sum_i x_i\lambda_i\pi_i}_{v(x)} + \sum_{i\in\{1,2\}} x_i \underbrace{\text{sgn}((\pi_{-i}-\pi_i)\lambda_i)}_{\kappa_i} \theta_i(|\boldsymbol{x}-\boldsymbol{z}|)$$

$$\theta_i(|\boldsymbol{x}-\boldsymbol{z}|) = \begin{cases} |(\pi_{-i}-\pi_i)\lambda_i| \, \frac{\sigma^2_{-i}}{\sigma^2_i+\sigma^2_{-i}} \mathbf{1}\{-i \in S\} & , i \in S \\ |(\pi_{-i}-\pi_i)\lambda_i| \, \frac{\sigma^2_{-i}}{\sigma^2_i+\sigma^2_{-i}} \mathbf{1}\{-i \notin S\} & , i \notin S \end{cases}$$

We can see that θ_i obeys dilution because it equals zero when attributes 1 and 2 have different status, and is weakly positive when they have the same status. □

Proof of Proposition 6. 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 have the same status in $|\boldsymbol{x}-\boldsymbol{z}|$, and do not have the same 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 have the same 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. We can write:

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

where $Z(\boldsymbol{x},\boldsymbol{z}) = \mathbf{1}\{i \in S^{(\boldsymbol{x},\boldsymbol{z})}\}\sum_{j\in S^{(\boldsymbol{x},\boldsymbol{z})}} \sigma_j^2 + \mathbf{1}\{i \notin S^{(\boldsymbol{x},\boldsymbol{z})}\}\sum_{j\notin S^{(\boldsymbol{x},\boldsymbol{z})}} \sigma_j^2$. The denominator is strictly positive. Since k has the same status as i in $(\boldsymbol{x},\boldsymbol{z})$ and not in $(\boldsymbol{x}',\boldsymbol{z}')$, $\sigma_k^2 \geq \sum_{i\neq k} \sigma_i^2$, implies $Z(\boldsymbol{x},\boldsymbol{z}) - Z(\boldsymbol{x}',\boldsymbol{z}') \geq 0$. Hence $\theta_i(\boldsymbol{x},\boldsymbol{z}) - \theta_i(\boldsymbol{x}',\boldsymbol{z}') \geq 0$.

Implicit associations. All attributes $j > 2$ have zero influence so trivially satisfy the assumption. Consider attribute $i \in \{1,2\}$. Since $k \in \{1,2\}$ as well, the influence function tells us that i has zero influence when i and k have different status, and weakly positive influence when they have the same status. □

A.4 Application to implicit risk and social preferences

Data access: For Exley (2016a) we use the replication data at `https://doi.org/10.1093/restud/rdv051`. For Ahumada et al. (2022) we use data kindly shared by the authors.
Data structure: Each experiment proceeds in three steps.

Normalization choice. For each participant, elicit using a choice list the smallest sure payment $\$X \in \{0, 2, \ldots, 30\}$ to a third party ("charity") that is preferred to \$10 for self.

Lotteries. Using X, construct a sequence of participant-specific simple lotteries. These pay out with probability $P \in \{0.05,\ 0.1,\ 0.25,\ 0.5,\ 0.75,\ 0.9,\ 0.95\}$ (Exley) or $P \in \{0.05,\ 0.25,\ 0.75,\ 0.95\}$ (Ahumada). Self lotteries, denoted by P^S, pay \$10 to self. Charity lotteries, denoted by P^C, pay $\$X$ to charity.

Choice lists. Elicit, using choice lists, preferences between each lottery and 21 different sure payoffs to self or to charity. We index these by $t = 0, ..., 20$. The sure payments are $Y_t^S = (0, 0.50, ..., 10)$ for self lotteries and $Y_t^C = (0, {}^{X}\!/_{20}, ..., X)$ for charity lotteries.

A bundle has three basic attributes: Recipient (Self/Charity), Prize, and Probability.

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. First, we represent the choice data in a space of two binary attributes: Social $\in$ {Selfish, Generous} and Risk $\in$ {Safe, Risky}. We construct the space so under Exley's null hypothesis we would expect the decision maker to be close to indifferent in all choices ("ambivalence", see Section 4). Second, we impute some choices that are not observed in the data.

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, Exley 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).

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 to Charity, and larger Prizes or Probabilities to smaller. So we cannot expect the participant to be close to indifferent ("ambivalence") in choice sets that vary on just one of these dimensions. Second, Prize and Probability are multivalued, while our framework requires binary attributes.

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. Ignoring wealth, the utility of a decision maker who satisfies standard risk preferences can be written as $v(c, y, p) = \pi(p)v\left(\frac{y}{1+\lambda c}\right)$. *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 (10).

$$v(c, y, p) = \pi(p)\left(\frac{y}{1 + \lambda c}\right)^\alpha. \tag{10}$$

Our approach amounts to selecting choices from the choice lists that can be described by two binary attributes where the decision maker is plausibly close to indifferent according to (10). We call them 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 such choices are not observed, so we analyze choices within each value of P.[52] 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 (10) would expect them to be close to indifferent. Consider the self lottery $(0, 10, P)$ that pays \$10 to Self with probability P. Equation (10) implies the following utilities are equal:

$$v\underbrace{(0, 10, P)}_{\text{Self lottery}} = v\underbrace{(1, (1+\lambda)10, P)}_{\text{Charity lottery}} = v\underbrace{\left(0, \pi(P)^{\frac{1}{\alpha}}10, 1\right)}_{\text{Self sure payoff}} = v\underbrace{\left(1, (1+\lambda)\pi(P)^{\frac{1}{\alpha}}10, 1\right)}_{\text{Charity sure payoff}} \tag{11}$$

Our approach will be to focus on choices defined by two 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

[52] We can think of this as multiple "slices" of a larger binary space with one attribute for each value of P.

the observer believes $L \approx 1+\lambda$ and $R(P) \approx \pi(P)^{\frac{1}{\alpha}}$.

We calibrate L using the initial normalization choice in the experiment: $L := {}^{X}/_{10}$ (which is the rate at which Exley compares self and charity payoffs). 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 implies that the participant slightly prefers any payoff LY to charity over Y to self, but is close to indifferent, consistent with the idea of "ambivalence."

We calibrate $R(P)$ using average behavior across participants, such that a participant with "average" risk aversion would be expected to be indifferent between lotteries and sure payoffs. Specifically, we compute the mean "switch point" in (1) choices between self lotteries/self sure payoffs and (2) charity lotteries/charity sure payoffs, measured as a fraction of the lottery prize, and set $R(P)$ equal to the mean of these two values. We do this separately by experiment and value of P. Thus, for probability P, the four two-attribute bundles are:[53]

$$\begin{aligned}
&(\text{Selfish, Risky}) &&= (0, 10, P) &&\text{(Self lottery)}\\
&(\text{Generous, Risky}) &&= (1, 10L, P) &&\text{(Charity lottery)}\\
&(\text{Selfish, Safe}) &&= (0, 10R(P), 1) &&\text{(Self sure payoff)}\\
&(\text{Generous, Safe}) &&= (1, 10LR(P), 1) &&\text{(Charity sure payoff)}
\end{aligned}$$

In any given choice list, we code the participant as choosing the lottery if their switch point exceeds the value implied by ambivalence, otherwise we code them as choosing the sure payoff. So, in a choice between (Selfish, Risky) and (Selfish, Safe) we code them as choosing the Risky bundle if they valued the lottery greater than $10R(P)$. In a choice between (Selfish, Risky) and (Generous, Safe) they choose the Risky bundle if they valued the lottery greater than $10LR(P)$.[54]

Imputing non-observed choices. To identify implicit Social preferences we need to observe choices where this attribute is non-shared while Risk is shared, but the data do not contain such choices. However, 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 choice (Generous, Safe) $\succ$ (Selfish, Safe).

Our calibration of the binary attribute space is constrained by the lotteries that we observe, whose prizes Exley also calibrated using 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).

[53] The CRRA assumption implies that (the log of) (10) 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.

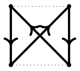
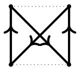
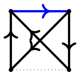
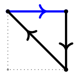
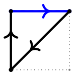
[54] Exley's analysis uses the midpoints between just-rejected and just-accepted payoffs to approximate certainty equivalents (i.e. points of indifference) of different lotteries. Our analysis uses the observed choices only, so is expressed in terms of strict preferences.

Table A1: Example datasets from from Exley sample

	Lottery win probability P							
Example	0.05	0.10	0.25	0.5	0.75	0.9	0.95	Type classification
1	–	–	–	–	(e)	(c)	(c)	Pro-Self only
2	(e)	(e)	(e)	(e)	(e)	(e)	(e)	Pro-Safe OR Self
3	–	(a)	–	(d)	(b)	(b)	(b)	Inconsistent
4	(b)	(e)	–	–	–	(e)	(e)	Pro-Safe only
5	–	(a)	–	(d)	(e)	(e)	–	Pro-Risky and Self

We show five participants from Exley's sample, selected to illustrate typical datasets that we observe. For each P we show the observed cycle type from Figure 6, if any, and the final column shows the participant's type classification.

Table A2: Frequencies of different cycles in Exley & Ahumada et al. datasets

Cycle			Exley	Exley	Ahumada	Pooled	Random
			7 probabilities	4 probabilities		4/7 probabilities	
Pro-Risky	(a)		0.025 (0.007)	0.026 (0.009)	0.018 (0.009)	0.023 (0.006)	0.063
Pro-Safe	(b)		0.078 (0.016)	0.081 (0.018)	0.045 (0.014)	0.069 (0.012)	0.063
Pro-Selfish	(c)		0.093 (0.017)	0.105 (0.021)	0.067 (0.016)	0.086 (0.013)	0.062
Pro-Self/Risky	(d)		0.143 (0.020)	0.119 (0.020)	0.103 (0.023)	0.132 (0.016)	0.125
Pro-Self/Safe	(e)		0.203 (0.026)	0.218 (0.029)	0.188 (0.029)	0.199 (0.021)	0.125
No cycle	(f)		0.458 (0.036)	0.451 (0.037)	0.580 (0.040)	0.492 (0.029)	0.563
Participants			86	86	56	142	

This table shows the frequency of each type of cycle in our analysis of data from Exley (2016) and Ahumada et al. (2022). Standard errors in parentheses are clustered at the participant level. First column shows results for all seven values of P in Exley's data. Columns 2 and 3 restrict to $P \in \{.05, .25, .75, .95\}$ for comparability. Column 4 uses all available data, and column 5 simulates random choice for seven and four probabilities respectively in proportion to study sample sizes. **Statistical tests.** Joint test of equality between Exley and Ahumada proportions (restricted to 4 probabilities): $p = 0.169$. Equality between rows in pooled dataset: $p(a = b = c) < .001$, $p(d = e) = 0.015$. Joint test versus random choice (pooled dataset): $p < .001$.

Table A3: Frequencies of different types in Exley & Ahumada et al. datasets, robust version

Type		Exley	Exley	Ahumada	Pooled	Random
		7 probabilities	4 probabilities		4/7 probabilities	
Inconsistent	(i)	0.023	0.000	0.000	0.014	0.084
		(0.016)	(.)	(.)	(0.010)	
Pro-Risky only	(ii)	0.035	0.023	0.018	0.028	0.073
		(0.020)	(0.016)	(0.018)	(0.014)	
Pro-Safe only	(iii)	0.105	0.116	0.089	0.099	0.073
		(0.033)	(0.035)	(0.038)	(0.025)	
Pro-Self only	(iv)	0.360	0.326	0.286	0.331	0.260
		(0.052)	(0.051)	(0.060)	(0.039)	
Pro-Risky and Self	(v)	0.035	0.023	0.036	0.035	0.145
		(0.020)	(0.016)	(0.025)	(0.015)	
Pro-Safe and Self	(vi)	0.070	0.058	0.054	0.063	0.145
		(0.027)	(0.025)	(0.030)	(0.020)	
Pro-Risky OR Self	(vii)	0.081	0.105	0.071	0.077	0.084
		(0.029)	(0.033)	(0.034)	(0.022)	
Pro-Safe OR Self	(viii)	0.128	0.186	0.232	0.169	0.084
		(0.036)	(0.042)	(0.056)	(0.031)	
No cycles	(ix)	0.163	0.163	0.214	0.183	0.052
		(0.040)	(0.040)	(0.055)	(0.032)	
Participants		86	86	56	142	

This table shows the classification of participants according to their revealed Preferences in our analysis of data from Exley (2016) and Ahumada et al. (2022). Standard errors in parentheses. First column shows results for all seven values of P in Exley's data. Columns 2 and 3 restrict to $P \in \{.05, .25, .75, .95\}$ for comparability. Column 4 uses all available data, and column 5 simulates random choice for seven and four probabilities respectively in proportion to study sample sizes. **Statistical tests.** Joint test of equality between Exley and Ahumada type distributions (restricted to 4 probabilities): $p = 0.965$. Equality between rows in pooled dataset: $p\big(ii = iii\big) = 0.017$, $p\big(v = vi\big) = 0.285$, $p\big(vii = viii\big) = 0.027$. Joint versus random choice (pooled dataset): $p < .001$. For robustness we require at least one preference in each cycle to be stricter than in the standard analysis. We operationalize this by requiring at least one lottery valuation to exceed the calibrated threshold by at least two choice list increments (rather than one as in the standard analysis).

A.5 Application to implicit racial discrimination

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]

Structural 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 all parameters (e.g., $\overline{v(\boldsymbol{x})}$ and the level of $\boldsymbol{\theta}$ are not separately identified), but we can identify changes such as $\overline{\kappa}_{race}\left(\theta^H_{race} - \theta^L_{race}\right)$, which are informative about the sign and quantitative importance of implicit preferences.

We illustrate by writing out expressions for the mean evaluations in the "Work Ethic concealed" treatment, to show what is identified. We set $x_{race} = -1$ for Black candidates. For the background attribute we use subscript b for brevity, and let $x_b = 1$ always. Recall that we do not observe the "Children" attribute in the data, so imposed from the start that its implicit preference equals zero.[55]

$$\begin{aligned}
\overline{y\,(\text{Black}, \text{White})} &= \overline{v\,(\text{Black})} - \overline{\kappa}_{race}\theta^L_{race} + \overline{\kappa}_b\theta^L_b \\
\overline{y\,(\text{Black}, \text{Black})} &= \overline{v\,(\text{Black})} - \overline{\kappa}_{race}\theta^H_{race} + \overline{\kappa}_b\theta^H_b \\
\overline{y\,(\text{White}, \text{Black})} &= \overline{v\,(\text{White})} + \overline{\kappa}_{race}\theta^L_{race} + \overline{\kappa}_b\theta^L_b \\
\overline{y\,(\text{White}, \text{White})} &= \overline{v\,(\text{White})} + \overline{\kappa}_{race}\theta^H_{race} + \overline{\kappa}_b\theta^H_b
\end{aligned}$$

Define $\beta_0 := \overline{v\,(\text{Black})} - \overline{\kappa}_{race}\theta^L_{race} + \overline{\kappa}_b\theta^L_b$, and $\beta_1 := \overline{v\,(\text{White})} + \overline{\kappa}_{race}\theta^L_{race} + \overline{\kappa}_b\theta^L_b$. We have:

$$\begin{aligned}
\overline{y\,(\text{Black}, \text{White})} &= \beta_0 \\
\overline{y\,(\text{Black}, \text{Black})} &= \beta_0 - \overline{\kappa}_{race}(\theta^H_{race} - \theta^L_{race}) + \overline{\kappa}_b(\theta^H_b - \theta^L_b) \\
\overline{y\,(\text{White}, \text{Black})} &= \beta_1 \\
\overline{y\,(\text{White}, \text{White})} &= \beta_1 + \overline{\kappa}_{race}(\theta^H_{race} - \theta^L_{race}) + \overline{\kappa}_b(\theta^H_b - \theta^L_b),
\end{aligned}$$

from which it is readily seen that $\overline{\kappa}_{race}(\theta^H_{race} - \theta^L_{race})$ and $\overline{\kappa}_b(\theta^H_b - \theta^L_b)$ are identified. We report estimates of 2× these quantities, which correspond to the relative increase in evaluations of candidates with $x_i = 1$, relative to those with $x_i = -1$, when changing from θ^L_i to θ^H_i.

[55] This restriction is only necessary because the Children attribute is not observed, it is not critical to our identification argument; if it was observed, we could also identify $\kappa_{children}(\theta^H_{children} - \theta^L_{children})$.