Cunningham, Tom; de Quidt, Jonathan

Working Paper

Implicit Preferences Inferred from Choice

CESifo Working Paper, No. 5704

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Cunningham, Tom; de Quidt, Jonathan (2016) : Implicit Preferences Inferred from Choice, CESifo Working Paper, No. 5704, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/128405

The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers

www.cesifo.org/wp

Implicit Preferences Inferred from Choice

Tom Cunningham
Jonathan de Quidt

CESIFO WORKING PAPER NO. 5704
CATEGORY 13: BEHAVIOURAL ECONOMICS
JANUARY 2016

ISSN 2364-1428

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 5704*

Implicit Preferences Inferred from Choice

Abstract

A longstanding distinction in psychology is between implicit and explicit preferences. Implicit preferences are ordinarily measured by observing non-choice data, such as response time. In this paper we introduce a method for inferring implicit preferences directly from choices. The necessary assumption is that implicit preferences toward an attribute (e.g. gender, race, sugar) have a stronger effect when the attribute is mixed with others, and so the decision becomes less "revealing" about one's preferences. We discuss reasons why preferences would have this property, advantages and disadvantages of this method relative to other measures of implicit preferences, and application to measuring implicit preferences in racial discrimination, self-control, and framing effects.

JEL-codes: D030, D830, J710.

Keywords: implicit discrimination, bias, judgement and decision making, choice-set effects.

Tom Cunningham
Facebook
tom.cunningham@gmail.com

Jonathan de Quidt
Institute for International Economic Studies
Stockholm / Sweden
jonathan.dequidt@iies.su.se

December 31, 2015

Preliminary draft, comments welcome.

Latest version available here: http://bit.ly/paper_implicit

Appendix available here: http://bit.ly/paper_implicit_appendix

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

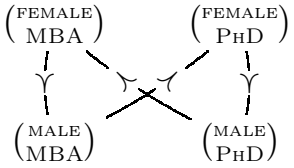
1 Introduction

In this paper we show how simple choices can, by themselves, reveal two separate sets of preferences. The idea is best illustrated with an intransitive cycle. Suppose you observe a recruiter's decisions between pairs of job applicants, each of whom is either male or female, and has either an MBA or a PhD. Suppose you observe that:

1. the recruiter chooses a female candidate over a male candidate whenever the two candidates' qualifications are the same,

2. the recruiter chooses a male candidate over a female candidate whenever the two candidates' qualifications differ.

Graphically, using $A \succ B$ to represent the choice of A from $\{A, B\}$:



These choices are inconsistent with maximization of a utility function. Nevertheless they form an intuitive pattern, which we describe as a "figure 8," and seem to reveal the existence of two distinct attitudes towards female candidates: a positive preference revealed in the vertical choice sets (between candidates who are otherwise identical), and a negative preference revealed in the diagonal choice sets (between candidates who differ in another respect besides gender).

Our paper generalizes this observation, that choices can sometimes reveal two distinct sets of preferences. We study choice over bundles of binary attributes (male/female, black/white, aisle/window), and we rank choice sets according to how *revealing* they are about each attribute. For example, in the diagram above, we say that the diagonal choice sets are less revealing about preferences over gender, compared to the vertical choice sets. We look for systematic differences in the preferences expressed in less revealing choice sets, and we define those preferences as *implicit* preferences (*explicit* preferences are those used in more revealing choice sets). Given the choices above we

would infer a positive explicit preference, but a negative implicit preference, for female candidates. The formal results in this paper are mainly concerned with deriving sufficient conditions on choice data from which we can establish that implicit and explicit preferences are different, regarding some attribute.

The formal framework we develop additionally shows how implicit preferences can be revealed in data on evaluations. For example, suppose we observe the sentences given by a judge to defendants, and we find that (1) when a black and a white defendant are sentenced alongside each other, there is no difference in the sentence received; but (2) when two black defendants are sentenced, they both get relatively long sentences, and when two white defendants are sentenced they both get short sentences. Under our model this behavior identifies an implicit preference in favor of white defendants.

We do not know of any prior theoretical papers which have identified this figure-8 pattern in choices, or which have shown how it can be used to identify implicit preferences; existing theories of menu-dependent preferences do not predict this pattern.[1] Nevertheless we think that the idea of implicit preferences being revealed by indirect choices taps into a commonsense understanding of decision-making, and most of our formal results correspond to natural intuitions. We discuss the few empirical papers we have found which can be interpreted as identifying implicit preferences.

Our introductory examples show how we can identify implicit discrimination - a topic of great recent interest.[2] But the possible applications are broad: in principle we can detect implicit preferences over any attribute, and there are many contexts in which we might expect them. Figure 1 shows a variety of figure-8 cycles in different domains. The choices indicated are our conjectures, to illustrate implicit preferences that we might expect to find.

- **Consumption.** Consider a person who chooses a diet soda over a full-sugar soda when they are of the same brand, but the full-sugar soda when they are of different brands. This reveals an explicit preference for diet soda, but an implicit

[1] E.g. "salience" (Bordalo et al. (2012)), "relative thinking" (Bushong et al. (2014)), "magnitude effects" (Cunningham (2012)), or "focusing" (Kőszegi and Szeidl (2011)). To the best of our knowledge, Cunningham (2014) is the only existing paper with an explicit identification of a figure-8 intransitive cycle.

[2] Bertrand et al. (2005) 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 giving a way of measuring the relative ambiguity of choices sets. Mullainathan (2015) gives a recent overview of evidence of implicit discrimination. People often make a distinction between statistical and taste-based discrimination: both are compatible with being implicit.

preference for full-sugar soda.

- **Self-other tradeoffs.** Consider a person who would always choose to give an object of value to charity, whether it is cash or goods. But when the payoffs are different (cash to one, goods to the other), then they choose in favor of themselves. This reveals an explicit preference in favor of the charity, but an implicit preference in favor of themself.[3]

- **Framing.** Consider a person choosing between two lotteries, one of which is described in terms of the probability of winning, the other in terms of the probability of losing. Suppose that the person is indifferent between two lotteries when they share the same objective payoffs, but when the lottery payoffs differ, they choose the one that is described with the emphasis on winning. These choices would reveal an implicit preference for the positive description, but no explicit preference. More generally, we think that many classical framing effects can be thought of as cases where a decision-maker has an implicit preference, and no explicit preference.[4]

- **Irrelevant Influences on Decision-Making.** Consider a person viewing two apartments, one on a sunny day and the other on a cloudy day. They are indifferent when the apartments are in the same building (i.e., identical), but when the apartments are in different buildings they tend to prefer the apartment viewed on the sunny day. This reveals an implicit preference for apartments viewed in good weather, but no explicit preference. [5]

Why would preferences change when the choice set becomes more revealing? We discuss a number of interpretations. Consider our introductory example of gender discrimination. First, people could be *unaware* of having a preference over gender, and

[3]The experiments in Exley (2015) have a similar structure, although that paper introduces a risk/safety tradeoff, rather than a cash/goods tradeoff, and does not use a figure-8 identification. We discuss Exley (2015) in detail later in the paper.

[4]Framing effects are usually defined as choice being influenced by a normatively irrelevant attribute, where "normative irrelevance" is imposed by assumption. Experiments rarely test whether framing effects survive in side-by-side evaluation, it is usually taken for granted that they do not. One exception is Mazar et al. (2013).

[5]A number of papers find that the weather influences economic decisions - Hirshleifer (2001), Simonsohn (2010), and Busse et al. (2013).

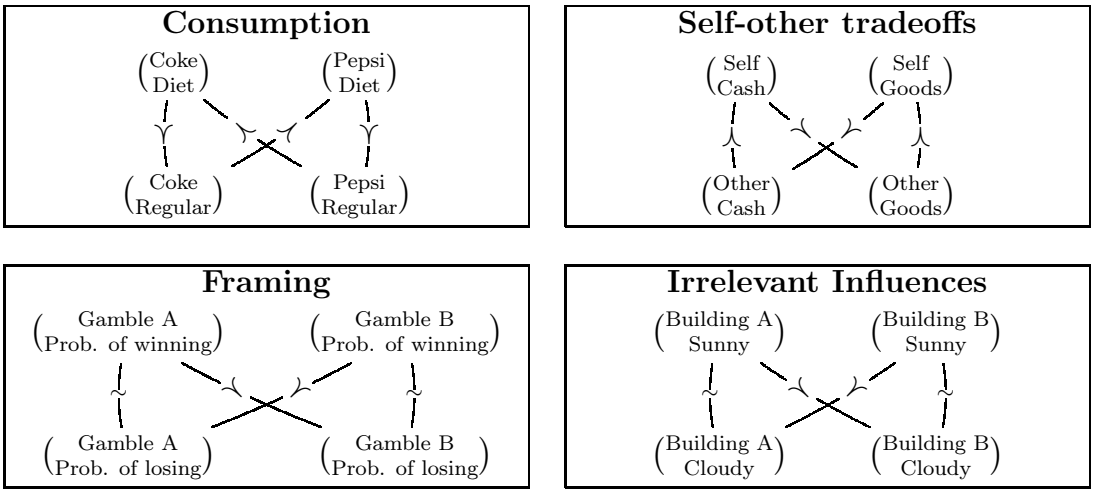


Figure 1: Figure 8 intransitivities applied to different domains.

correct for it only insofar as they can detect it in their own instincts, implying that gender would have a bigger effect on judgments in less revealing choice sets.[6] Second, people could be *aware* of their gender bias, but would like to conceal it from an observer, i.e. they wish to signal their preferences (the decision-maker could be their own observer, as in models of self-signaling). Again this implies that gender would have a larger effect in less revealing choices. Third, people could be constrained by what we term *ceteris paribus* rules. Such a rule could be, for example, "never choose a man over an equally-qualified woman." Most of our theoretical work is agnostic about the underlying cause of implicit preferences, but we also discuss ways in which the interpretations can be distinguished - most simply, someone who signals their preferences will be constrained by precedents set by prior choices, while someone who learns their preferences (as in the unconscious influences model) will not.

Economic implications of implicit preferences. There are many influential theories of human behaviour in which motives are in some sense hidden: Freudian and subsequent psychoanalytic theory; recent claims in social psychology about unconscious

[6]Suppose you get a good feeling about the male candidate, and a bad feeling about the female candidate. If they have the same qualifications, then you can infer exactly why you have different feelings. If they have different qualifications, then the feeling may be attributed, in part, to the difference in qualifications. We formalize this theory in the Appendix using an application of the model in Cunningham (2014). In that model the conscious system must rely on pre-conscious systems for interpreting information, and therefore can be influenced by aspects of a stimulus that it regards as normatively irrelevant.

processes;[7] claims in judgment and decision-making about unconscious influences;[8] evolutionary theories of self-deception in humans and other animals;[9] and economic theories of signaling in social behaviour.[10] There is also a case to be made from introspection: we often are unsure about, for example, whether we would have liked a wine equally much if it had been $10 instead of $50; whether we would have liked an academic paper as much if it had been submitted under a different name; whether we would have treated a student the same way if they had been of a different gender or race. This ignorance leaves opens a door for implicit influences. More narrowly, as economists we are interested in decisions concerning race, charitable giving, politics, and status goods, and in each case it is widely thought that people have serious internal conflicts in making these decisions. Our paper gives a rigorous foundation for estimating the strength of implicit influences in all of these domains.

Our theory has implications for applied industrial organization. It predicts that demand for a good can vary systematically with features of the choice set. Firms who are aware of the effects we analyze will bundle implicitly desired features or products along with other features to make the purchase less revealing, for example by bundling pornographic pictures with journalism.[11]

Our measure of implicit preferences can be compared with the Implicit Association Test, which uses response time in a categorization task.[12] An important advantage of our measure is that it is based only on ordinary decision-making, so needs little additional interpretation to be used in interpreting economic outcomes, and can be computed directly from observational data.

Prior experiments on implicit preferences. A few prior experimental studies have relied on the intuition that we are attempting to formalize: Snyder et al. (1979) on implicit discrimination against the handicapped, Exley (2015) on implicit preferences over giving to charity, and Bohnet et al. (2015) on implicit gender discrimination. For each of these papers we show that, although they study implicit preferences in our sense, the statistical tests which they use to identify implicit discrimination are imperfect (i.e., they would identify implicit preferences where none exist), and we describe alternative

[7] For example, in the recent popular books "Blink", "Subliminal", "The Hidden Brain", "The Invisible Gorilla", and "Incognito".

[8] e.g. Kahneman (2011).

[9] Von Hippel and Trivers (2011)

[10] Spence (1973), Hanson (2008)

[11] Chance and Norton (2009).

[12] Greenwald et al. (1998)

appropriate tests. We also reanalyze an existing dataset from DeSante (2013) and find evidence for an implicit preference in favor of white over black welfare applicants.

The next section contains the main formal results. We first state assumptions under which implicit preferences can be inferred from (1) a three-element intransitive cycle; (2) a four-element intransitive cycle ("figure 8"); (3) evaluation data where the same outcome is evaluated differently alongside different comparators; and (4) a pair of such comparator-sensitive evaluations. Section 3 discusses alternative ways of identifying implicit preferences; how to analyze different types of dataset; plausible foundations that generate implicit preferences; and relates our interpretations to existing literature. Section 4 discusses four existing empirical papers. Section 5 gives a brief overview of economic applications, and Section 6 concludes. Three Appendices contain proofs, a statement of the models that generate implicit preferences (rule-following, signaling, and implicit knowledge), and additional formal results.

2 Model

The paper has many formal results, so we begin by providing a summary. We consider *outcomes* which are bundles of binary attributes (e.g., male/female, short/tall, day/night). In most of the paper we consider data on either choice between a pair of outcomes, or continuous evaluation of both members a pair of outcomes. We derive parallel techniques for detecting implicit preferences in each type of dataset. Most of our results establish conditions under which the data are sufficient to establish the direction of an implicit preference, i.e. whether the implicit preference is positive or negative with respect to some attribute. The identification is entirely through observing violations of rationality - either by observing an intransitive cycle, or by observing that evaluation of an outcome changes when the identity of the other outcome being evaluated changes (the "comparator").

If we impose the restriction that implicit preferences can exist over only one attribute (for example just over gender), then the task is relatively straight-forward: we can infer the direction of the implicit preference by observing either a single 3-element intransitive cycle in choices, or a single comparator-effect on evaluation. The task becomes more complicated when implicit preferences could exist over multiple attributes, for example, over both gender and qualification. Much of the formal work shows how such effects can be disentangled.

For each result we have tried to present a minimal set of assumptions. This comes at a cost of greater complexity. A reader who is less interested in the details can skip most of the discussion of assumptions.

Results for choice data:

1. **Right-triangle cycle.** Observing an intransitive cycle among three outcomes, where one outcome is *between* the other two (defined below), reveals an implicit preference for at least one of the attributes on which the outcomes differ. If sufficiently many right-triangle cycles are observed, implicit preferences over a single attribute can be inferred.

2. **Figure-8 cycle.** Observing a figure-8 intransitive cycle (as in the introduction) reveals an unambiguous implicit preference for one attribute.

Additional results for choice data:

3. **Isosceles cycle in a ternary space.** In some settings it is natural to consider attributes with three values (e.g. male/female/no gender). Under a minor extension to our definition of betweenness, to cover the ternary attribute space, we can identify an unambiguous implicit preference from a single cycle with three elements (an *isosceles* cycle).

4. **Aggregation.** We give conditions under which aggregate choice data (i.e., between-subjects data) can establish an implicit preference.

Results for evaluation data:

5. **Scissor effect.** Observing that evaluation of some outcome changes when its comparator changes, in a manner that satisfies betweenness, reveals a disjunction among a set of implicit preferences over the outcome's attributes.

6. **Parallel scissors.** Observing that the evaluations of a pair of outcomes, which differ only in one attribute, move in opposite directions when there are symmetric changes to each of their comparators, reveals an unambiguous implicit preference over that attribute.

Additional results for evaluation data:

7. **Joint and separate evaluation.** Observing that evaluations of a pair of outcomes, which differ only in one attribute, move in opposite directions when moving from separate to joint evaluation, reveals an unambiguous implicit preference over that attribute.

Theoretical foundations for implicit preferences. We outline in the paper, and formally present in an Appendix a few natural foundations which generate implicit preferences.

8. **Linear implicit preferences.** We first introduce a general model, called linear implicit preferences, in which all the binary attribute space results hold (i.e. all results above except number 3).

9. ***Ceteris paribus* rules.** A decision-maker constrained by what we term *ceteris paribus* rules will exhibit linear implicit preferences.

10. **Signaling.** A linear-Gaussian model of a decision-maker who wishes to signal his preferences to an observer will exhibit linear implicit preferences.

11. **Implicit knowledge.** A linear-Gaussian two-system decision-maker, with imperfect knowledge of their own preferences, will exhibit linear implicit preferences in choice, provided the outcomes in the choice set differ by no more than two attributes.

2.1 Choice in a Binary Space

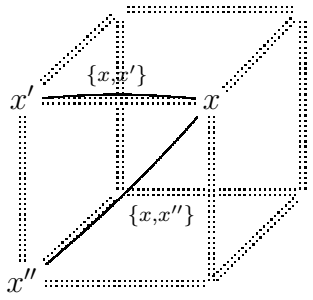
The space of outcomes is defined by n binary attributes, i.e. $X = \{0,1\}^n$. In many cases we will without loss of generality consider outcomes with $x_i = 1$, $\forall i$. We consider only binary choice sets for the majority of this paper, so the set of choice sets is $\mathcal{A} = \{\{x,y\} : x,y \in X, x \neq y\}$. A menu-dependent utility function is a function $u : X, \mathcal{A} \to \mathbb{R}$.

We assume that choice sets can be ranked in terms of *revealingness* regarding each attribute. Formally we assume that there exists a set of simple orders among choice sets, denoted $\geq_i$, where $A \geq_i B$ means that choice-set A is weakly more revealing than choice-set B with respect to attribute i.[13] The interpretation of revealingness differs between our foundations. However they all satisfy the following assumption: that a

[13]The symbols $>_i$, $=_i$, and $\neq_i$ are defined in the usual way relative to $\geq_i$.

choice set is less revealing about an attribute when more other attributes are bundled with it - in other words, when it becomes more diluted.

To state this clearly we first define an outcome x' as being *between* x and x'' if it is a convex combination: in the following diagram x' is between x and x''. [14] The betweenness assumption implies that the pair $\{x, x'\}$ is relatively more revealing about the horizontal dimensions.



Definition 1. For any $x, x', x'' \in X$, x' is **between** x and x'' if for all i, either $x'_i = x_i$ or $x'_i = x''_i$.

We now make the assumption that a strict increase in the dimension-wise distance between two elements will lower the revealingness about the attributes on which they already differ.

Assumption 1 (Betweenness)**.** *For any $x, x', x'' \in X$, if x' is between x and x'' then $\{x, x'\} \geq_i \{x, x''\}$ for all i such that $x_i \neq x'_i$.*

We now define implicit preferences: roughly speaking, a decision-maker has a positive implicit preference over an attribute if they become more likely to choose outcomes with that attribute when revealingness with respect to that attribute decreases. As an example:

Example 1. Consider a decision-maker with a positive implicit preference for whites, and consider a white and a black candidate who are equal in every other respect. If the white candidate is preferred in one context, then they will also be preferred when revealingness with respect to race decreases.

[14] The diagram is drawn in 3 dimensions, but can represent an arbitrary number of attributes bundled into three groups: the attributes on which x and x' disagree plotted on the horizontal axis, those on which x' and x'' disagree plotted in the vertical axis, and those on which all three elements agree plotted in the remaining axis.

Definition 2. We say that $u(x, A)$ has **relative implicit preferences** $\lambda \in \{-1, 0, 1\}^n$ with respect to a set of orderings on $\mathcal{A}$, $\{>_i\}_{i=1}^n$, if, for any $x, x' \in X$, and $A, B \in \mathcal{A}$, normalizing $x_j = 1, \forall j$, such that for every i with $x'_i = 0$,

$$\begin{aligned} A \geq_i B &\Leftrightarrow \lambda_i \geq 0 \\ A \leq_i B &\Leftrightarrow \lambda_i \leq 0, \end{aligned}$$

then

$$u(x, A) > u(x', A) \implies u(x, B) > u(x', B).$$

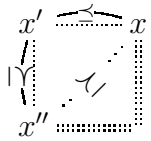
The vector λ summarizes the implicit preferences: if $\lambda_i = +1$, then we say that u has a positive implicit preference for attribute i, if $\lambda_i = -1$, negative implicit preferences, and if $\lambda_i = 0$, no implicit preference. Our definition assumes separability of implicit preferences: for example, if the attributes are male/female and white/black, then we allow for an implicit preference for men, and for white candidates, but not one which applies just to white men. However we make no assumption about the separability of $u(x, A)$ in x_i and x_j. Specifically, conditional on the choice set we allow an arbitrary ranking of outcomes, but changes to that ranking must obey the vector of implicit preferences when revealingness changes.[15]

This definition of implicit preferences, along with betweenness, is sufficient to make basic inferences from certain intransitive choices. We will use $\succeq$ as a shorthand to denote choice from a binary choice set, i.e. $x \succeq x'$ if and only if $x \in c(\{x, x'\})$, which in turn is true if and only if $u(x, \{x, x'\}) \geq u(x', \{x, x'\})$.

We first show that a 3-element intransitive cycle which satisfies betweenness establishes a disjunction among implicit preferences. In the following diagram the observed choices reveal that the decision-maker must have a negative implicit preference for one of the attributes which x and x'' disagree on, because the relative preference for x over

[15]Consider a set of candidates who are White (W) or Black (B) and Male (M) or Female (F), and a decision-maker with a positive implicit preference for males and none over race. For any given choice set A we allow the decision-maker's preferences to be non-separable in race and gender, for example: $u(WM, A) = u(BF, A) > u(BM, A) = u(WF, A)$. However, if a choice set B is less revealing with respect to gender, the decision-maker's preferences must shift in favor of males, i.e. $u(WM, B) \geq u(BF, B)$ and $u(BM, B) \geq u(WF, B)$.

x'' declines in the less revealing comparison (the hypotenuse of the triangle).



Proposition 1. *[right triangle cycle]For any $x, x', x'' \in X$, if x' is between x and x'', and $x \succeq x' \succeq x'' \succeq x$, with at least one relation strict, then u must have a negative implicit preference for one of the attributes on which x and x'' differ (normalizing $x_i = 1$, $\forall i$).*

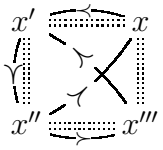
Observing a single right-triangle cycle only establishes a disjunction among implicit preferences. If we are willing to assume that there exists an implicit preference on at most one, given attribute, a single right-triangle cycle is sufficient to identify it.

Example 2. Suppose we observe the following preferences over cellphones: $\binom{\text{iPhone}}{\text{Gold}} \succ \binom{\text{Android}}{\text{Silver}} \succ \binom{\text{Android}}{\text{Gold}} \succ \binom{\text{iPhone}}{\text{Gold}}$. Under the maintained assumption that implicit preferences can exist only over brand, this is sufficient to identify a positive implicit preference for iPhones.

When we allow for implicit preferences on multiple dimensions, observing a set of such cycles can identify the existence of an unambiguous implicit preference over a single attribute. Define the **span** m of a right-triangle-cycle as the number of dimensions on which the outcomes that lie on the hypotenuse differ (i.e., $m = \sum_{i=1}^{n} 1\{x_i \neq x_i''\}$).

Proposition 2. *To establish an unambiguous implicit preference from right-triangle-cycles of span m requires observing at least 2^{m-1} such cycles.*

Importantly, note that when outcomes differ in at most two attributes (such as our Male/Female-MBA/PhD example), only *two* right-triangles are needed. For example, consider the following:



However, note that, at least when the cycles span only two attributes, to identify an implicit preference we must observe either (i) two preferences on the horizontal dimension which go in different directions (a *non-monotonicity*); or (ii) two indifferences

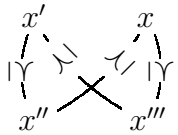
along the horizontal dimension. Further discussion and examples of combining multiple cycles can be found in the Appendix.

There is a more parsimonious way of inferring implicit preferences: from a figure-8 cycle of intransitivities. This requires an additional assumption: that revealingness depends only on the set of dimensions which differ between the outcomes considered. For example, we assume that the two choice sets, $\left\{\binom{\text{MALE}}{\text{MBA}}, \binom{\text{FEMALE}}{\text{PHD}}\right\}$ and $\left\{\binom{\text{FEMALE}}{\text{MBA}}, \binom{\text{MALE}}{\text{PHD}}\right\}$ are equally revealing about each of the attributes. In the Appendix we show that this assumption will hold in all the underlying models of implicit preferences that we consider.[16]

Assumption 2 (Equivalence)**.** *For any $x, x', x'', x''' \in \mathcal{A}$, if, for all $i \in \{1, \ldots, n\}$, $|x_i - x'_i| = |x''_i - x'''_i|$ then for all $i \in \{1, \ldots, n\}$, $\{x, x'\} =_i \{x'', x'''\}$.*

Example 3. Consider the diagram in Proposition 3. The following pairs of choice sets are equally revealing about both attributes: $\{\{x, x'\}, \{x'', x'''\}\}$, $\{\{x', x''\}, \{x, x'''\}\}$, and $\{\{x', x'''\}, \{x, x''\}\}$.

Proposition 3 (figure 8 cycle)**.** *Suppose Equivalence holds. For any $x, x', x'', x''' \in X$ (normalizing $x_i = 1, \forall i$), if (1) x' is between x and x'', (2) $x'''_i \neq x''_i \iff x_i \neq x'_i$, and (3) preferences are such that:*



with at least one preference strict, then u must have a negative implicit preference for an attribute on which x and x''' differ.

Note that unlike the use of two right-triangles, we do not require a non-monotonicity or indifference along one dimension.

2.2 Evaluation in a Binary Space

We now turn to data on evaluations, applicable to, for example, bids in an auction, statements of willingness to pay, assignment of scores in judging sports, etc.. $\mathcal{A}$ now represents the set of *evaluation sets*: pairs of outcomes to which the decision-maker

[16] If equivalence did not hold then a figure-8 could occur without any North-South implicit preferences, e.g. suppose that $\{x, x''\}$ was more revealing about East-West preferences than $\{x', x'''\}$, and caused an increase in East-West sensitivity such that $x'' \succ x$.

simultaneously assigns evaluations. A menu-dependent evaluation function is a function $y : X, \mathcal{A} \to \mathbb{R}$.[17]

The main results in this section parallel those in the section on choice. We first slightly strengthen the betweenness assumption:

Assumption 3 (Strong betweenness)**.** *For any $x, x', x'' \in X$, if x' is between x and x'' then $\{x, x'\} \geq_i \{x, x''\}$ for all i such that $x_i \neq x'_i$, and $\{x, x'\} \leq_i \{x, x''\}$ for all i such that $x_i = x'_i$.*

This extends the prior assumption of betweenness by specifying that revealingness weakly *decreases* for common attributes within a pair when we increase the number of attributes which differ. Strong betweenness holds in all of our foundational models, and follows from the logic of signal extraction: if we think of the evaluation of each outcome in the evaluation set as informative about the value associated with the common attributes, then reducing the correlation of those evaluations will increase the accuracy of our inference about the common attributes.

Second, we appropriately modify the definition of implicit preferences. Previously, in less revealing situations, preferences would switch in favor of the implicitly favored outcome, all else equal. We now assume that the effect is not just marginal but absolute: in less revealing situations the evaluations given to the implicitly favored outcome will increase and the evaluation of the disfavored outcome will decrease. Formally, the evaluation of an outcome x will increase with a change in evaluation set if the new set is more revealing about attributes that implicitly disfavor x, and less revealing about attributes that implicitly favor x.

Definition 3. We say that $y(x, A)$ is an **implicit evaluation function** with $\lambda \in \{-1, 0, 1\}^n$, if, for any $x \in X$ and $A, B \in \mathcal{A}$ such that (normalizing $x_j = 1, \forall j$):

$$\begin{aligned} A \geq_i B &\Leftrightarrow \lambda_i \geq 0 \\ A \leq_i B &\Leftrightarrow \lambda_i \leq 0, \end{aligned}$$

then

$$y(x, A) \leq y(x, B).$$

[17] We emphasize the necessity of evaluation *sets*. It is not possible to extract implicit preferences solely from evaluations made in isolation.

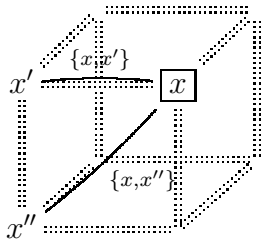
From this it follows that, if we observe the evaluation of x increase when the comparator shifts strictly away from x,[18] then there must exist either a negative implicit preference for one of x's attributes which the original comparator agreed on (for which revealingness has increased), or a positive implicit preference for one of x's attributes which the original comparator disagreed on (for which revealingness has decreased). We call this a "scissor effect."

Proposition 4 (Scissor effect)**.** *For any $x, x', x'' \in X$, with x' between x and x'' and*

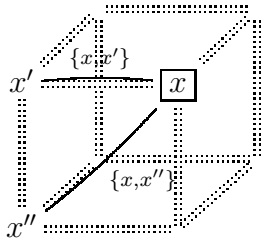
$$y(x, \{x, x''\}) > y(x, \{x, x'\}),$$

then (normalizing $x_i = 1\ \forall i$) either (i) y has a positive implicit preference for some attribute i ($\lambda_i > 0$) on which x and x' disagree ($x_i \neq x'_i$); or (ii) y has a negative implicit preference ($\lambda_i < 0$) for some attribute i on which x and x' agree ($x_i = x'_i$).

Example 4. Consider the diagram below. If the evaluation of x increases with a change of comparator from x' to x'' this implies either a positive implicit preference for x's value on the horizontal dimension (for which revealingness has decreased) or a negative implicit preference for x's value on another dimension (for which revealingness has increased).



Proof. Consider the case graphically:



[18]Where "strictly away" is used in the betweenness sense, that the old comparator is between x and the new comparator.

Suppose, for the sake of contradiction, that y has a weakly negative implicit preference for every attribute on which x and x' disagree (for which $\{x, x''\} <_i \{x, x'\}$), and a weakly positive implicit preference for every attribute on which x and x' agree (for which $\{x, x''\} >_i \{x, x'\}$). Then, by the definition of implicit preferences, it must be the case that $y(x, \{x, x''\}) \leq y(x, \{x, x'\})$, contradicting the premise. □

We discuss in Appendix 8.3 how the disjunctions derived from multiple scissor effects can be combined to infer unambiguous implicit preferences. As before, a single scissor will be sufficient to identify a unique implicit preference if we are willing to assume that there are no implicit preferences over other attributes.

Example 5. Suppose we observe the following pattern of willingness-to-pay for cellphones:

$$y\left(\begin{pmatrix}\text{iPhone}\\ \text{Gold}\end{pmatrix}, \left\{\begin{pmatrix}\text{iPhone}\\ \text{Gold}\end{pmatrix}, \begin{pmatrix}\text{Android}\\ \text{Silver}\end{pmatrix}\right\}\right) > y\left(\begin{pmatrix}\text{iPhone}\\ \text{Gold}\end{pmatrix}, \left\{\begin{pmatrix}\text{iPhone}\\ \text{Gold}\end{pmatrix}, \begin{pmatrix}\text{Android}\\ \text{Gold}\end{pmatrix}\right\}\right).$$

Under the maintained assumption that implicit preferences can exist only over brand, this is sufficient to identify a positive implicit preference for iPhones.

There exists a more parsimonious way of identifying unambiguous implicit preferences . Suppose we observe two different scissor effects composed of outcomes that are identical but flipped with respect to attribute i, meaning that the second scissors is composed of outcomes identical to the first, except that they have the opposite value of attribute i, compared to the corresponding outcomes in the first scissors. This essentially enables us to focus on the influence of attribute i, "controlling for" the influence of the attributes $j \neq i$. If the two scissors cause opposite shifts in evaluation, we identify an unambiguous implicit preference over attribute i. We term this a *parallel scissor effect*.

The parallel scissor effect relies on two assumptions. First, the equivalence assumption, described above, so that revealingness is comparable between the evaluation sets.[19] Second, we assume that implicit preferences are *monotonic*, in the following sense:

[19]The equivalence assumption is stronger when applied to evaluation than when applied to choice. Briefly - equivalence could be violated in a signaling model if there is differential uncertainty about the weights on each of a pair of attributes - e.g. if your evaluation of $\binom{\text{BLACK}}{\text{PhD}}$ and of $\binom{\text{WHITE}}{\text{PhD}}$ could be differentially revealing about your PhD-preference, if an observer is more certain of your white-preference than your black-preference. This issue does not seem to be important in choice, where an observer only gets information about the difference between the two realizations of an attribute (i.e., the black-white difference).

Assumption 4 (Monotonicity)**.** *For any $x, x' \in X$ and $A, B \in \mathcal{A}$, normalizing $x_i = 1, \forall i$, if, for all j with $x'_j = 0$,*

$$\begin{aligned} A \geq_i B &\Leftrightarrow \lambda_i \geq 0 \\ A \leq_i B &\Leftrightarrow \lambda_i \leq 0, \end{aligned}$$

then,

$$y(x', A) < y(x', B) \implies y(x, A) < y(x, B).$$

Example 6. Consider a decision-maker with a positive implicit preference for males, and a male and female candidate who are identical in other respects. If the evaluation of the female increases when switching from evaluation set A to B, then the evaluation of the male must also increase.

Monotonicity is not in fact guaranteed by our basic definition of implicit preferences, because switching from A to B can change revealingness about other attributes as well as gender. Monotonicity imposes that the effect of these other attributes on evaluation cannot overwhelm the effect of gender.

Proposition 5 (Parallel scissor effects)**.** *For some i, and $\underline{x}, \bar{x}, \underline{x}', \bar{x}', \underline{x}'', \bar{x}'' \in X$, with $\bar{x}_i = \underline{x}_i + 1$, $\bar{x}_j = \underline{x}_j, \forall j \neq i$, $\bar{x}'$ between $\bar{x}$ and $\bar{x}''$, and $|\bar{x} - \bar{x}'| = |\underline{x} - \underline{x}'|$, and $|\bar{x} - \bar{x}''| = |\underline{x} - \underline{x}''|$, if we observe*

$$\begin{aligned} y(\bar{x}, \{\bar{x}, \bar{x}'\}) &\geq y(\bar{x}, \{\bar{x}, \bar{x}''\}) \\ y(\underline{x}, \{\underline{x}, \underline{x}'\}) &\leq y(\underline{x}, \{\underline{x}, \underline{x}''\}), \end{aligned}$$

with one inequality strict, then $\lambda_i > 0$.

Proof. First note that, by equivalence, just two evaluation functions are invoked, denote them

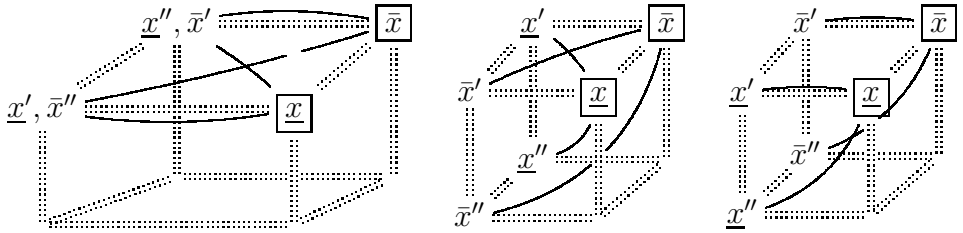
$$\begin{aligned} y^A(\cdot) &= y(\cdot, \{\bar{x}, \bar{x}'\}) = y(\cdot, \{\underline{x}, \underline{x}'\}) \\ y^B(\cdot) &= y(\cdot, \{\bar{x}, \bar{x}''\}) = y(\cdot, \{\underline{x}, \underline{x}''\}), \end{aligned}$$

from which we can rewrite the inequalities,

$$\begin{aligned} y^A(\bar{x}) &\geq y^B(\bar{x}) \\ y^A(\underline{x}) &\leq y^B(\underline{x}). \end{aligned}$$

Assume that $\lambda_i \leq 0$. But, by the monotonicity assumption, this implies B is weakly more favorable to $\bar{x}$ than $\underline{x}$, contradicting the two observed inequalities (assuming one is strict). □

This proposition yields a surprisingly rich variety of tests for implicit preferences. These tests can be put into three categories, depending on whether $\bar{x}'$ and $\bar{x}''$ agree with $\bar{x}$ on the attribute of interest. Consider the following three diagrams, which illustrate the three types of parallel scissor effects, constructed around the outcomes $\bar{x}$ and $\underline{x}$ which differ on attribute i (e.g., gender). We normalize $\bar{x}_j = 1, \forall j$, hence $\bar{x}_i = 1$ (male) and $\underline{x}_i = 0$ (female). If the evaluations of $\bar{x}$ and $\underline{x}$ shift in opposite directions when their comparators undergo equivalent transformations, this must be due to an implicit preference over attribute i.



Type 1 Type 2 Type 3

Type 1 $\bar{x}'$ agrees with $\bar{x}$ on attribute i, and $\bar{x}''$ disagrees. Thus the shift from $\bar{x}'$ to $\bar{x}''$ increases revealingness about attribute i (as does the shift from $\underline{x}'$ to $\underline{x}''$ when evaluating $\underline{x}$). Then, a positive implicit preference for males would be revealed if we observe both that (a) evaluation of the female candidate $\underline{x}$ increases when her comparator changes from female to male, and (b) evaluation of the male candidate $\bar{x}$ decreases when his (symmetric) comparator changes from male to female.

Type 2 both $\bar{x}'$ and $\bar{x}''$ disagree with $\bar{x}$ on attribute i. Then the shift from $\bar{x}'$ to $\bar{x}''$ decreases revealingness about attribute i (as does the shift from $\underline{x}'$ to $\underline{x}''$ when evaluating $\underline{x}$). Then, a positive implicit preference for males would be revealed if we observe both that (a) evaluation of the female candidate $\underline{x}$ increases when

her comparator becomes more similar on the non-gender dimensions, and (b) evaluation of the male candidate $\bar{x}$ decreases when his (symmetric) comparator becomes more similar.

Type 3 both $\bar{x}'$ and $\bar{x}''$ agree with $\bar{x}$ on attribute i. Then the shift from $\bar{x}'$ to $\bar{x}''$ increases revealingness about attribute i (as does the shift from $\underline{x}'$ to $\underline{x}''$ when evaluating $\underline{x}$).Then, a positive implicit preference for males would be revealed if we observe both that (a) evaluation of the female candidate $\underline{x}$ increases when her comparator becomes more similar on the non-gender dimensions, and (b) evaluation of the male candidate $\bar{x}$ decreases when his (symmetric) comparator becomes more similar. This third case may be the weakest method of detecting implicit preferences, because the change in revealingness might be expected to be small, given that there is no variation in the attribute of interest (attribute i) within either of the evaluation sets.

2.2.1 Joint and Separate Evaluation

With a minor additional assumption, the same logic will also allow us to infer implicit preferences from comparison of *joint* and *separate evaluation* of two outcomes $\bar{x}$ and $\underline{x}$, where joint evaluation considers evaluation set $\{\bar{x}, \underline{x}\}$, while separate evaluation considers $\{\bar{x}\}$ or $\{\underline{x}\}$. The assumption is that evaluation of a single outcome is equally revealing as evaluation of a pair of identical outcomes ($\{x\} =_i \{x, x\}, \forall i$), implying that $y(x|\{x\}) = y(x|\{x, x\})$.[20] Then, if we find that $\underline{x}$ and $\bar{x}$ move in opposite directions when evaluated jointly relative to when evaluated separately, this reveals an implicit preference with respect to attribute i.

Example 7. Consider a female and male candidate who are identical on all other attributes. If the female candidate's evaluation increases and the male candidate's evaluation decreases when evaluated jointly, we identify an implicit preference for male candidates.

2.2.2 Testing Evaluation Data

Given these results, how should one analyze a dataset on evaluations? Suppose there are 2 attributes, implying 4 outcomes and 16 conditional evaluations of the form

[20]Note that x is trivially between x and any x'.

$y(x|\{x, x'\})$.[21] Then, for each attribute, we can run 10 separate tests for implicit preferences.[22] Each test could identify either a positive implicit preference, a negative implicit preference, or an ambiguous result. A test of the theory as a whole can be performed by checking that the data never identifies, for the same attribute, both positive and negative implicit preferences.

3 Foundations of implicit preference

We discuss three formal models which would generate implicit preferences. Derivations are given in an Appendix.

3.1 Signaling

Suppose that you are concerned about the outward appearance of your preferences, for instance you might tend to prefer unhealthy snacks, but be embarrassed about it. A choice set that is more revealing about attribute i will tend to be one for which the observer's beliefs about your preferences over i are more sensitive to your choice. The more sensitive the observer's beliefs, the more the decision-maker will attempt to disguise their true motivations, generating implicit preferences. We give a fully worked-out model in the Appendix: a decision-maker possesses, for each attribute, a coefficient representing their intrinsic preference, and a coefficient representing their concern about how other people perceive their intrinsic preferences. The sign of the second coefficient corresponds to the direction of the implicit preference that can be identified from choice.[23]

Some economists have argued that much social behavior is motivated by signaling concerns, for example that education is to signal ability Spence (1973), conspicuous con-

[21] If there are n attributes then there are 2^n possible outcomes, and so 2^{2n} potential observations of the form $y(x|x')$. For each attribute there will be 2^{n-1} pairs of outcomes, $\bar{x}$ and $\underline{x}$, and then a variety of $\bar{x}'$ and $\bar{x}''$. Our calculations include evaluation sets with the same element repeated twice (technically a multiset).

[22] There are two possible choices for $\bar{x}$. If $\bar{x}' = \bar{x}$ then there are 3 choices for $\bar{x}''$. There are two other choices for $\bar{x}'$, and for each $\bar{x}''$ is unique (it is the exact opposite of $\bar{x}$). Thus there are ten tests in total.

[23] The model in the Appendix assumes that the observer has independent Gaussian priors over the intrinsic preferences. We also assume that the observer's priors are mean-zero, and explain why betweenness can be violated when this is not true. We assume a naive observer, i.e., the observer does not appreciate that the decision-maker has signaling motivation, but we believe that similar results would obtain with a sophisticated observer.

sumption is used to signal status Veblen (1899), or generosity is used to signal altruism Bénabou and Tirole (2006).[24] If correct then demand for education, consumption and generosity should be lower in less revealing choice situations.

The signaling model can also be interpreted as self-signaling, as in Benabou and Tirole (2003) and Bodner and Prelec (2003), in which you distort your actions to persuade your own future self that you are generous, or clever, or hard-working. In these models, for the signal to be effective, the future self must be assumed to forget the present-self's motivations or circumstances.

3.2 Maximizing with *Ceteris Paribus* Rules

Implicit preferences could be generated by an ordinary decision-maker who is constrained by one or more rules, each of which requires that a certain attribute be preferred when all other attributes are equal. We call these *ceteris paribus* rules, and give a formal model of this type of decision-making in the Appendix. Each rule will manifest as an implicit preference, and therefore can be identified from behavior using the conditions we have derived.

This type of decision-making appears in a variety of real-world contexts: in a bureaucracy, rules are often explicitly written as ceteris paribus rules, e.g. "never appoint a male when there is an equally qualified female candidate."[25] Universities are often forbidden from discriminating on the basis of race (and are often thought to discriminate on attributes correlated with race). It seems that many people take care to never *overtly* discriminate on the basis of race, sex, or political affiliation, but do allow those factors do influence their decisions when the comparison is less revealing. In individual decision-making we sometimes observe people following rules such as "you must always choose the diet version of a soda, when available."[26]

Viewed from the perspective of signaling these rules express an "innocent until proven guilty" philosophy, under which people are only penalized when their action incontrovertibly reveals a forbidden preference. This behavior is difficult to reconcile

[24] See Hanson (2008) for an expansive argument about the importance of signaling.

[25] Or "fly economy class when it is available," or "if two bids are otherwise equivalent, choose the lowest bidder."

[26] It has commonly been observed that people adopt "personal rules": inflexible principles, often interpreted as means of self-control. For example: going to the gym at the same time every day; never making a withdrawal from your savings account; always forgoing dessert. Models which rationalize personal rules include Ainslie (1992), Bénabou and Tirole (2004), Bodner and Prelec (2003), Brocas and Carrillo (2008).

theory	typical evidence	typical findings
Freudian "deep psychology"	dreams, slips of the tongue, forgetting, jokes	sexual fixations
1970s social psychology[28]	influence of primes on judgment and decision-making	self-serving bias, social desirability bias
implicit motives[29]	Thematic Apperception Test (free response to a picture)	desire for power, achievement, emotional affiliation
implicit associations[30]	response time in an association task	discriminatory associations

Table 1: Some Theories of Subconscious/Implicit Motives

with the linear-Gaussian signaling model, in which the expression of implicit preferences varies continuously with revealingness.[27]

Finally, *ceteris paribus* decision-making is a special case of decision by "lexicographic semiorder", discussed in the Appendix.

3.3 Implicit Knowledge

The idea that there are important subconscious influences on behavior did not become widespread until the 19th century (Ellenberger (1970)). Since then there have been many theories of such influences, and various techniques of identifying them. A few are summarized in Table 1 below. All of these techniques remain controversial. We consider our method to be an alternative means of identifying unconscious influences on behavior: a factor is unconscious if its influence judgment systematically differs with the revealingness of the situation.

For example, suppose we find that judgment of a drink's flavor is influenced by its color; judgment of a person's honesty is influenced by the clothes they wear; judgment of the value of a house is influenced by the glossiness of the brochure; or judgment of the severity of a crime is influenced by whether it was committed by a Republican or Democrat. Each influence could be conscious or unconscious: we can test for the consciousness of each of these influences by seeing if they vary as the revealingness is varied - e.g., by eliciting judgments side by side.

[27]Under the linear-Gaussian model, even when evaluating a man and woman side by side, who are otherwise equal, they would not receive the same evaluation: the intrinsic preferences and signaling preferences will be traded off, meaning any bias would be diminished, but not eliminated.

In an Appendix we state a model with two stages: you first get an "intuition" about the value of each outcome, and then you adjust each intuition, based on additional considerations, before making a final decision. Formally, two mental processes work sequentially, each forming an estimate of value, but each with access to private information. This implies that you have intuitions that are informative because they incorporate knowledge to which you do not have conscious access. This model predicts systematic *comparison* effects in decision-making, because each new element in the choice set can reveal different information about the implicit knowledge. The model in this paper is a simplified version of that given in Cunningham (2014).

We show that the model meets our definition of implicit preferences in choice when the outcomes differ in at most two respects. A positive implicit preference for an attribute, e.g. for male job candidates, implies one of two things: (1) that the decision-maker believes gender to be irrelevant, but has unconscious positive associations with men; (2) that the decision-maker does believe gender to be relevant, but has unconscious negative associations with men (and hence the difference in evaluation declines more in revealing choice sets). When the outcomes differ in more than 2 attributes then the techniques we use (triangle and figure-8s) are not appropriate for identifying implicit knowledge in this model. We discuss this further in the Appendix.[31]

We believe that this model can give a good account of framing effects: they are due to associations that are *normally* relevant, but irrelevant in the current context. This corresponds to a common informal description of biases being byproducts of rational heuristics (Tversky and Kahneman (1981)). We give examples and further discussion in section 6.3.

A final variation on the implicit knowledge model would be one with *motivated* bias: we sometimes talk of people deceiving themselves into making decisions by finding an excuse for their preferred outcome.[32]

[31]The model could also explain implicit race or sex bias under the assumptions that (1) people have learned associations with race or sex that they are unaware of (or they are unaware of their magnitude), and (2) people think that those associations are irrelevant for typical decisions. This explanation is similar to common descriptions of behavior in the implicit association test (IAT) - that people are unaware of their race-based instincts, and attempt to correct for them. However if this explanation is correct it remains a puzzle why people remain unaware of their associations despite relatively frequent experience with making race-based and sex-based decisions.

[32]It would be possible to write down a model with an expert and a decision-maker, such that the expert's bias will be mixed into their advice, and derive a prediction that the expert's preferences will manifest as implicit preferences. However it is much easier to achieve this pattern in decisions if the decision-maker is imperfectly informed about the expert's biases, otherwise the decision-maker could simply correct the advice to account for their bias. Thus there remains an element of this self-deception

3.4 Distinguishing Between Interpretations

The interpretations given above cannot be distinguished on the basis of simple binary choice or evaluation, because they all fit the general model of implicit preferences. However we discuss a variety of ways to distinguish between them, with: (1) a change in incentives or observability of the choice, (2) variation in the preceding choice set, (3) variation in the order of preceding choice sets, or (4) choice from larger choice sets.

First, under the "signaling" interpretation the decision-maker will be sensitive to the implementation of their decision: the strength of implicit preferences should therefore be increasing in the probability of the decision being implemented (because this decreases the relative importance of the signaling motive), and decreasing in the probability of the decision being observed (which increases the relative importance of the signaling motive). Under an "implicit knowledge" interpretation neither change should affect the relative weight of implicit and explicit preferences.

Second, the models have different implications about the effects of preceding choice sets. Under implicit knowledge if some choice set is completely revealing about attribute i then the decision-maker will learn their preference over i, and so this should eliminate implicit preferences over i in subsequent choices. For example, if I am asked to choose between $\binom{\text{MALE}}{\text{MBA}}$ and $\binom{\text{FEMALE}}{\text{MBA}}$, this will reveal to me my implicit bias, and I should not exhibit any implicit preferences over gender in subsequent questions, for example in tradeoffs between MALE/FEMALE and OXFORD/CAMBRIDGE. This is not true in the signaling model.[33]

Third, the *ceteris paribus* model implies that choices will set precedents, and so constrain subsequent choices, leading to *order* effects that would not occur in the implicit knowledge model. Consider the following two sequences of three choice sets, which are identical except for the order of the first two sets:

$$\left(\left\{\binom{\text{FEMALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\},\left\{\binom{\text{FEMALE}}{\text{PHD}},\binom{\text{MALE}}{\text{MBA}}\right\},\left\{\binom{\text{MALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\}\right)$$
$$\left(\left\{\binom{\text{FEMALE}}{\text{PHD}},\binom{\text{MALE}}{\text{MBA}}\right\},\left\{\binom{\text{FEMALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\},\left\{\binom{\text{MALE}}{\text{MBA}},\binom{\text{MALE}}{\text{PHD}}\right\}\right)$$

A decision-maker with implicit knowledge will condition on the information learnt in

that is unexplained, because it seems that most people are aware of the direction of their own biases – e.g., in favor of their preferred political party, in favor of unhealthy foods, against physical exertion – yet those biases still seem to distort their judgments.

[33]This point courtesy of Luke Miner.

the prior choice sets, but the order of those choice sets should not matter. In contrast a *ceteris paribus* decision-maker who is not allowed to choose a male over a female, and who chooses the male candidate in the first choice set, will be forced to choose, in the third choice set, whichever candidate has the qualification which the male had in the first set. This follows from assuming that they are forbidden from making a choice which, when combined with prior choices, implies a violation of a *ceteris paribus* constraint through transitivity.[34]

Finally, the models differ in their predictions about choice from 3-element choice sets. Consider the following two choice sets:

$$\binom{\text{FEMALE}}{\text{MBA}} \qquad\qquad \binom{\text{FEMALE}}{\text{PHD}}$$

$$\binom{\text{MALE}}{\text{MBA}} \quad \binom{\text{MALE}}{\text{PHD}} \qquad\qquad \binom{\text{MALE}}{\text{MBA}} \quad \binom{\text{MALE}}{\text{PHD}}$$

A *ceteris-paribus* decision-maker with a rule not to choose a man over a similar woman, and a sufficiently strong implicit preference for men, would choose $\binom{\text{MALE}}{\text{PHD}}$ from the left-hand choice-set, and $\binom{\text{MALE}}{\text{MBA}}$ from the right-hand one, a violation of GARP. A decision-maker with implicit knowledge would never make such choices because both choice sets would be equally informative about her unknown preference parameters, and so would both evoke the same set of preferences (at least, using the Appendix's linear-Gaussian version of implicit knowledge).

4 Discussion

4.1 Comparison of Alternative Ways of Identifying Implicit Preferences

Our theoretical exposition takes as given that we know what the decision-maker would choose or what her evaluation would be in each choice or evaluation set. In practice, of course, choices and evaluations must be observed or elicited, opening up a number of

[34]The two sequences are chosen so that, by the third step, the history is identical, but the order varies. A similar effect of precedents seems natural in the signaling model, but it is somewhat more difficult to model the desire for consistency. Interestingly, the order effects generated by *ceteris paribus* decision-making allow for strategic effects in agenda-setting: the decision-maker's final choice can be manipulated by gradually revealing choices, and eliciting intermediate choices.

interesting methodological issues.

4.1.1 Choices in a Binary Space

History Effects in Within-Subject Studies. If we do not know what a given subject will choose from each choice set, the natural approach is to measure it by presenting her with all relevant choices and recording what she does, i.e. collect within-subject data. However, there are reasons to expect significant history effects: one's decision is influenced by the prior choice set (in the "implicit knowledge" story), or by prior choices (under the other foundations). This makes within-subject data more complicated to interpret.[35] Under some assumptions, separating choices with decoy questions or time intervals (if she is forgetful), or making her feel un-observed by exploiting administrative data (if she has a signaling motive) can alleviate the problem.

Heterogeneity in Between-Subject Studies. If we instead use between-subject data then we have the problem that significant between-subject heterogeneity of preferences could again make implicit preferences undetectable, no matter how well-calibrated is the choice set. To establish the existence of at least one decision-maker with intransitive preferences over outcomes a, b, c the aggregate choices must violate the triangle inequality: for a cycle of 3 elements the average choice probability must exceed $\frac{2}{3}$, (i.e., $P(a \succ b) + P(b \succ c) + P(c \succ a) > 2$).[36] Heterogeneity in preferences among subjects will tend to push choice probabilities towards $\frac{1}{2}$, implying that moderate heterogeneity in explicit preferences could make it impossible to prove the existence of implicit preferences.

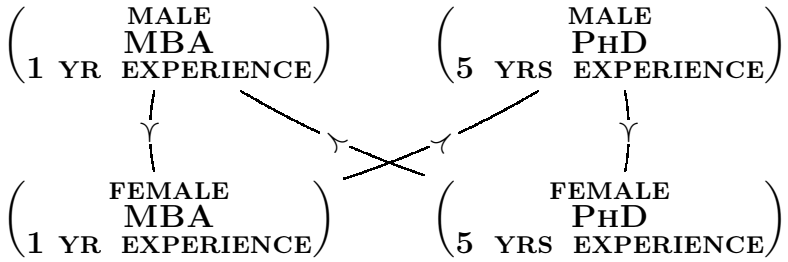
Fortunately, a straightforward way to increase our ability to identify implicit preferences is to collect data on indifference (indeed, we noted the value of observed indifference in the theory). For some attributes we expect there to be little heterogeneity

[35] This challenge is not unique to our proposed approach (not to mention the further challenge of ensuring multiple choices are incentive-compatible), and is the reason why between-subject designs are more commonly used in economics and psychology experiments.

[36] For intuition, note that if $\frac{2}{3}$ of subjects report $a \succ b$, $\frac{2}{3}$ report $b \succ c$ and $\frac{2}{3}$ report $c \succ a$ this could be rationalized by a subject pool in which $\frac{1}{3}$ of subjects have transitive preference $a \succ b \succ c$, $\frac{1}{3}$ $b \succ c \succ a$ and $\frac{1}{3}$ $c \succ a \succ b$). If the cycle has four elements the requirements are stronger: the average choice probability must be greater than $\frac{3}{4}$ (Regenwetter et al. (2011)). To *statistically* establish cyclical preferences in a finite sample will require still higher fractions because of sampling variation. The problem of heterogeneity is reflected in the observation that, although there are many well documented and strong framing effects, there are few clear demonstrations of intransitive choices in the laboratory (Regenwetter et al. (2011)).

in direct comparisons: for example, we might expect that close to 100% of subjects are indifferent about gender in direct comparisons. Then *any* difference in choice ratios in indirect comparisons is sufficient to identify an implicit preference.[37]

Calibration. Even a decision-maker with substantial implicit preferences will not reveal them if we do not observe the right kind of choices. For example, if the decision-maker has a significant preference for MBAs over PhDs, then the choice-sets presented in the introduction might not detect any implicit preference over gender, even if one exists, because they will always choose the candidate with the MBA. This is essentially a calibration problem. Fortunately, by varying an additional attribute we may be able to bring the decision-maker closer to indifference:[38]



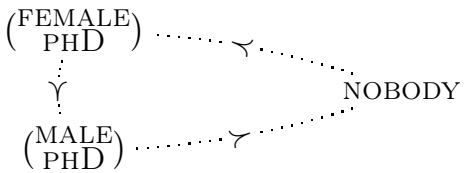
4.1.2 Choices in a Ternary Space (Isosceles cycles).

Some choices do not naturally fit into a binary space. Suppose we observe a recruiter who would hire a female PhD over a male PhD, and hire a male PhD over hiring nobody,

[37] If all subjects are indifferent between male and female candidates in direct comparison then rationality implies that the preferences between MBA and PhD should be identical, irrespective of associated gender, and any departure from this pattern would reveal an implicit preference.

[38] A common way to deal with such calibration problems is by using "multiple price list" to find the indifference point between two bundles of goods, e.g. answering "what value of x would make you indifferent between $\binom{1 \text{ can spinach}}{3 \text{ cans corn}}$ and $\binom{x \text{ cans spinach}}{1 \text{ can corn}}$?" This is sometimes called "matching." A disadvantage is that the act of choosing an x could be psychologically different than making a binary choice, and so have less external validity when predicting choice behavior. Also note that here we are treating "1 year experience" and "5 years experience" as two poles of a binary attribute.

but would also hire nobody over hiring a female PhD, i.e. an intransitive cycle:



These choices seem to reveal an unambiguous implicit preference for male over female employees, yet do not have a natural analysis in a space composed only of binary attributes. We discuss in an Appendix how the binary model can be extended to license such an inference from cycles like the above, which we call *isosceles* cycles. An isosceles cycle is more parsimonious than a figure-8 cycle, having only three outcomes. In many cases it may also be more sensitive (inducing more variation in revealingness). For example, the choice set $\{(\text{FEMALE}\ \text{PHD}), \text{NOBODY}\}$ seems intuitively less revealing about gender preferences than is the choice set $\{(\text{FEMALE}\ \text{PHD}), (\text{MALE}\ \text{MBA})\}$).

Nevertheless in this paper we principally concentrate on outcomes which can be represented in a binary space, for a number of reasons: (1) identifying a set of binary attributes in a set of outcomes often is less controversial; (2) for each isosceles cycle that identifies an implicit preference, a corresponding figure-8 cycle can often be constructed;[39] (3) isosceles cycles could occur for reasons other than the existence of implicit preferences. For example if people are sensitive to the range of outcomes in a choice set, as in the theory of Hsee and Zhang (2010) who argues that people tend to be more sensitive to an attribute when there is more variation in that attribute. As we argue below, a figure-8 cycle is difficult to explain with existing theories of decision-making, and so is clearer evidence of implicit preferences.

4.1.3 Evaluation.

Using data from evaluation, instead of choice, will tend to be more sensitive to implicit preferences for three reasons.

Variation in Revealingness. Evaluations allow for greater variation in revealingness. This is because we can measure implicit preferences over an attribute using data on evaluations of choice sets which include only one realization of an attribute, for example by comparing evaluations among groups that are men-only, women-only, and

[39] In the example above, by replacing the "nobody" outcome with candidates who have MBAs.

mixed, whereas inference from choice can identify implicit gender preferences only from mixed sets.

Calibration. Second, with evaluation calibration problems largely disappear, i.e. the method can detect even very subtle implicit preferences, while, as noted above, choice data can only detect implicit preferences that are large enough to change the ranking of outcomes.

Power. Third, evaluation can be continuous, rather than discrete, tending to increase statistical power.

Disadvantages. However a disadvantage of evaluation is that it may be less natural in domains where choice is more common, and therefore findings regarding evaluation would have lower external validity. Additionally, a choice is explicitly comparative, forcing subjects to consider every element of the choice set. Instead, when forming an evaluation, subjects do not have to consider every element of the evaluation set, yet will reveal their implicit preferences only if they do so. As a result one might wrongly conclude that a decision-maker has no implicit preferences, where instead they simply performed their evaluations separately.

Heterogeneity. Suppose we observe average evaluations over a population–as would occur in a between-subjects experiment–how does this affect our analysis? In particular, if we treat the average evaluations as those of a representative agent, and infer the implicit preferences of that agent, what can we then conclude about the population? If the direction of implicit preferences are not aligned within the population (i.e., if some people have a strictly positive preference for attribute i, and others have a strictly negative one), then a representative agent may not exist (i.e., there may be no single set of implicit preferences which rationalize the average evaluations). However we conjecture that if implicit preferences are aligned in this sense, then a representative agent will exist, and thus the population's implicit preferences can be identified with the implicit preferences of that agent.

4.1.4 Sequential Evaluations

We often observe people making evaluations in a series: e.g., giving ratings to a series of job applicants. If we are willing to assume that the evaluation set consists of the

current outcome under consideration, plus the most recently considered outcome, then it is straightforward to apply our existing results for evaluation. We provide more details in Appendix 8.4.

4.1.5 Other Issues

In the Appendix we discuss the relationship with other types of cycles: equilateral cycles, and cycles which indicate non-separable implicit preferences (section 11.1), and extension to larger choice sets (section 11.2).

4.2 Competing Theories

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

Contingent weighting. Models of contingent weighting in multi-attribute choice, like our theory, assume that preferences depend on the choice set.[40] However existing theories rely on a very different intuition: they assume that the sensitivity to a given attribute depends on the observed distribution over that attribute. For example sensitivity to race would depend on the distribution of black and white elements. However in our model sensitivity to race will instead depend on the distribution of the *other* attributes - e.g., a decision-maker with implicit racial preferences would become more sensitive to race when the distribution of other attributes such as education becomes more dispersed. None of the recent contingent-weighting models is consistent with a figure-8 intransitivity.[41]

[40] For example in Kőszegi and Szeidl (2011) sensitivity is positively related to the range of values on an attribute, in Bushong et al. (2014) it is negatively related to the range, in Cunningham (2012) it is negatively related to the average, and in Bordalo et al. (2012) it is - roughly - negatively related to the proportional range (range divided by the average).

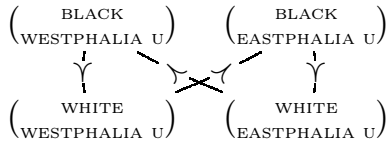
[41] Formally, suppose the utility function is separable in each attribute, in the sense that it can be written as,

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

where a_i^j is the ith attribute of the jth element of the choice set, A, then a figure-8 intransitivity could never occur because - using the gender example - the marginal distribution of the gender attribute

A similar point applies to the literature on comparing joint and separate evaluation of outcomes: Hsee et al. (1999) give many examples. Most of these studies find that people are more sensitive to an attribute when presented jointly - for example the difference in WTP for high-quality and low-quality goods tends to be higher in joint evaluation. Hsee et al. (1999) argue that this increased sensitivity is a general feature of joint evaluation, called "evaluability".[42] Again, this is a quite different principle to that used in implicit preferences. This mechanism could generate isosceles intransitivities and joint-separate differences in evaluation. However it could not generate a figure-8 cycle, by an analogous argument to footnote 41. See Cunningham (2012) for a Bayesian rationalization of increased sensitivity in joint evaluation.

Inference from the choice set. We have assumed that the attributes of an outcome are not informative about the value of other outcomes in the choice set. If they were informative then inference from the choice set could in principle rationalize *any* pattern in choice. The relevant question is what types of prior beliefs could generate the patterns we observe, and whether those beliefs seem realistic. Suppose we observe a cycle in choice among job candidates who vary in both race and in the school at which they studied:



These decisions could be rationalized by a decision-maker who (1) believes black candidates are better than white candidates, all else equal; but (2) believes that white candidates typically go to better schools, and therefore infers the quality of the school from the choice set. Thus in the diagonal choice sets they will prefer white candidates not because they are white, but because they went to the school that white people go

remains the same in all four choice sets, thus the difference in attribute-utility (u_i) between "Male" and "Female" must remain the same. The two diagonal choice-sets must evoke the same utility function, because they have the same marginal distributions, and that utility function prefers Male to Female, all else equal. But this contradicts the choice observed in the vertical choice sets (where Female is chosen over Male). Separability holds for all the models discussed above except Bordalo et al. (2012), but that model cannot generate intransitive cycles in binary choices with two attributes.

[42]For example subjects were found to state a higher WTP for a dictionary with 10,000 entries when it was evaluated alone, than when it was evaluated alongside a dictionary with 20,000 entries and a torn cover. Kahneman and Frederick (2005) discuss a similar phenomenon: that subjects are generally more sensitive to changes in within-subjects experiments than in between-subjects experiments. The theory is further developed in Hsee and Zhang (2010)

to. In practice we believe that this alternative explanation of implicit preferences is not a realistic concern in most of our applications because (1) most examples we discuss use familiar attributes, so the scope for learning from the choice set seems small; and (2) the explanation requires that the *intrinsic* value of an attribute be opposite to its *informational* value (in this case, being white is a negative signal about the person, but it is a positive signal about the things which covary with being white).[43]

Inattention / Heuristics. Because much of our identification comes from comparing simple to complex choices (or direct to indirect choices), we may worry that inconsistencies are due to variation in complexity, as in models of inattention (Sims (2003), Caplin and Martin (2011), Woodford (2012)). It is intuitive that a decision-maker could become less sensitive to an attribute in a more complex choice situation, however we have not been able to find an inattention model in which an increase in complexity causes the polarity of an attribute to *reverse*, as necessary for the figure 8.[44]

4.3 Other Measures of Implicit Preference

We discuss a number of measures. An influential paper, Dana et al. (2007), reports a variety of experiments which show that pro-social choices are affected by "wiggle room." Each of their experiments falls under a different heading in the classification that follows.

Rationalization. Cherepanov et al. (2013) (CFS) propose a model of "rationalization" which is related to ours. Agents possess both a *true* preference relation and a set of *rationalizable* preference relations. A decision-maker will choose the item which is her favorite among those that would be chosen by at least one of the rationalizable

[43]To explain a figure-8 with indifferences on the vertical comparisons, the intrinsic value of the vertical attribute must be zero and the informational value be non-zero, for example if race is believed to have no value in itself, but white students tend to go to better colleges.

There are cases where informational effects are certainly important: e.g., suppose one attribute is "Old Grouse" vs "Johnny Walker", and the other attribute is "labelled as Whisky of the Year" vs "no label." Naturally a decision-maker is indifferent about which bottle has the label, when the bottles are of the same brand, but strictly prefers the bottle with the label when they are of different brands. In general we presume that attributes are informative about the *token*, not the *type* of an outcome.

[44]As was the case with inference, a figure-8 with indifferences could come from inattention if sensitivity to an attribute goes to zero in simple choices; though we are not aware of an inattention model with this feature.

preferences.[45]

The CFS model has a similar spirit to our model: their "rationalizable" preferences roughly corresponds to our "explicit" preference.[46] There are two important differences: (1) while CFS study choice among atomic elements, we study choice among bundles of attributes, making it easier to extrapolate behavior to new situations under our approach (for example detecting an implicit racial preference among one set of candidates has implications for choices among a completely different set). (2) We allow for choice to be a continuous mixture of implicit and explicit preferences, while in CFS the effects are binary: either a choice is rationalizable or not (the *ceteris paribus* model shares that feature with CFS).

Adding Noise (list elicitation and random response). Some experiments measure how preferences vary when noise is added to a decision. In the "list elicitation" method subjects are given a set of statements and record just the number of statements that they agree with.[47] In the "random response" method subjects are given one statement, and then flip a coin with the instructions to mark "yes" if either (a) the coin lands heads (unobserved by the experimenter), or (b) they agree with the statement.

Under a signaling model these experiments could help identify implicit preferences - loosely reasoning that noise lowers the revealingness of a decision - so these techniques should reveal implicit preferences when compared with responses to the same questions asked separately.

A problem with both of these techniques is that, although adding noise reduces the incentive to distort, at the same time it increases the *ability* to distort, because the noise is private information to the decision-maker, allowing the decision-maker

[45]The principal working example is the following vignette: *"Dee decides to take time off from work to see a movie. However, prior to leaving the office she is informed that a colleague is in the local hospital and can accept visitors that afternoon. Dee reconsiders her decision to go to the movie and, instead, stays at work."* Dee's choices violate the weak axiom (WARP). Under the CFS model we can infer two facts: (1) Dee has the "true" preference order,

$$\text{MOVIE} \succ \text{WORK} \succ \text{VISIT SICK FRIEND},$$

but that (2) none of Dee's "rationalizable" preferences rank MOVIE above VISIT SICK FRIEND.

[46]In addition our *ceteris paribus* model, when there is a single ceteris paribus rule, obeys WWARP, the Weak Weak Axiom of Revealed Preference, the axiom which characterizes the CFS model (or, more generally, a lexicographic semiorder, discussed in Manzini and Mariotti (2012)).

[47]Miller (1984) the technique is also called "item count" or "randomized response." A post on Andrew Gelman's blog (Gelman, 2014) surveys some empirical work with these techniques and gives a pessimistic summary of their usefulness.

to misreport the noise. This means that adding noise has an ambiguous effect on reporting. This has been found in the data: for example, John et al. (2013) found that the random-response method did not increase the fraction of people who admitted to an embarrassing statement (in this case, admitting having cheated on an earlier test), in fact it decreased the number who admitted to it. John et al. conjecture that this was because some subjects answered "no" even when the coin landed heads-up (when they should have answered "yes", if following the instructions) due to a strong desire to signal that they did not cheat.[48] Thus either an increase *or* a decrease in reporting under these protocols can be interpreted as evidence for under-reporting in the ordinary protocol.

One solution to this problem is to add noise only after subjects make a decision, instead of letting subjects to add the noise themselves. This is used by Dana et al. (2007): they found that when decision-makers faced a chance of a donation decision not being implemented (and their decision was not observed by the beneficiary, only the implementation) then they tended to make more selfish decisions.

Verbal Explanation of Decisions. A series of papers has used verbal explanations of the decision-process as the dependent variable in a manipulation. Subjects are first asked to make a decision between two outcomes (bundles of attributes), and then asked what factors were most important in their decision. Papers in this literature typically report finding that (a) some attribute affects the decision without being described as important, while (b) whichever other attribute is *correlated* with the first attribute is described as important. For example Hodson et al. (2002) find that, in choice among black and white college applicants, subjects reported being uninfluenced by race, but when the white applicant had better grades then subjects were more likely to rate grades as an important factor.[49]

These studies are clearly related to the method advocated in this paper, but differ in using verbal judgments rather than choices. For instance, under a signaling inter-

[48]The same logic holds for the item-count technique: when asked to sum the statements that they agree with, subjects have an increased ability to distort their answers. Gelman (2014) mentions some experiments that find this perverse effect.

[49]Interestingly Norton, Vandello & Darley (2004) use the same technique and find the opposite effect - a pro-black bias - perhaps because of difference in subject pools. Norton, Vandello and Darley (2004) find that, in a choice between candidates for a job in construction, when the female candidate had less education, then subjects were more likely to rate education as important. Norton (2010) found that, in a choice between magazines, when the magazine with swimsuit photos also had articles on sport, then subjects were more likely to rate sports-coverage as important.

pretation the decision-maker is reporting the weights they put on attributes directly rather than weights being inferred by an observer.

Choice over Choice Sets. A variety of studies find situations in which subjects strictly prefer smaller choice sets (i.e., they will pay to avoid being given an additional alternative). In Dana et al. (2006) and Lazear et al. (2012) subjects have the choice whether to play a dictator game, or opt out of it at some cost, and many choose to opt out. Andreoni et al. (2011) similarly find that people are willing to pay to avoid a charity collector. These have a natural signaling interpretation: the decision-maker prefers to leave money on the table, than to make a selfish choice that is observed by the recipient. In our signaling framework we identify concern for reputation via changes in choices or evaluations between more or less revealing situations, while here it is identified by willingness to pay to avoid a revealing situation.

Signalling and Crowding Out. In Benabou and Tirole (2003) providing an incentive for an action can change the signaling value of that action. In particular they predict a u-shaped effect: incentives decrease the signaling value when the action is rare (or unexpected), and increase the signaling value when the action is common (or expected). This occurs when the observer's priors are single-peaked - implying that an action is least informative about one's preferences when the observer puts a 50% chance on you performing the action (informativeness here means the difference in posterior means). They thus predict that providing an incentive for a rare pro-social act (e.g. giving blood) can crowd out the signaling incentive, because it causes the act to become less diagnostic about one's pro-sociality.

Their results are related to the results from our signaling model: both show how changing the bundling of attributes can change the signaling value of a choice. They consider adding a feature with a known positive value, i.e. an incentive. Our model deals with adding features that have unknown values (with mean-zero expected value). We therefore consider their approach to be complementary.[50]

Choice of information. A variety of biases seem to be identified by choice to be strategically *ignorant*. A good example is reported in Dana et al. (2007)'s "hidden information" experiment. They find that subjects' choices are sensitive to the payoffs

[50]Bodner and Prelec (2003) also have a self-signaling model. Mijović-Prelec and Prelec (2010) has a useful discussion on the difference between self-deception and merely having biased beliefs.

of their partner (a standard finding), but that, in addition, subjects prefer to remain ignorant about how their partner's payoff depends on the choice; and that when subjects are ignorant they tend to make the choice which maximizes their own payoff.[51]

Dana et al. refer to an "illusory preference for fairness." We might say that the possibility of not revealing the payoffs of the partner makes the decision under the treatment "less revealing," though the example does not fit neatly into our binary attribute framework. Their result is striking in particular because choosing to reveal should make the decision maker weakly better off (she is better able to trade off fairness and efficiency if she knows the payoffs), and strictly so unless she is very selfish. An interpretation which relates revealingness to the number of steps of reasoning required to determine if an action was selfish or not seems intuitively appropriate here.

Rabin (1995) proposes that people often treat moral considerations not as ends in themselves, but as constraints on maximizing consumption. This motivation can be identified in information-seeking behavior: such people will choose to avoid information whenever that information will, in expectation, lead to decisions that lowers their selfish utility.[52]

Automatic Responses. Nosek et al. (2011) survey experimental measures of implicit social cognition. Most of those measures ask subjects to perform a classification task quickly, and test whether classification speed or accuracy is affected by semantic relationships among the stimuli used. Most famous is the Implicit Association Test, but there are many other variants.

[51]Subjects choose between allocations of money, denoted (self,other). Control subjects had to choose between a fair allocation $(5,5)$ and an unfair allocation $(6,1)$. Treatment subjects were given a choice between $(5,X)$ and $(6,Y)$. Pressing a button would reveal X and Y, which were either equal to 5 and 1, or 1 and 5 respectively. The generic pattern of choices was to choose $(5,5)$ under the control, and $(not\,reveal,(6,Y))$ under the treatment, consistent with the uncertainty giving some "moral wiggle room."

[52]For example I might sincerely believe that the suffering of animals is not sufficient to become a vegetarian, but also avoid learning more for fear that I might revise upwards my estimate of suffering, and be forced to stop eating meat. This theory will only have empirical bite if the selfish payoff is nonlinear in beliefs (e.g., if my decision to eat meat is all-or-nothing). A more general treatment of this could identify, from choices over distributions of information (as in Kamenica and Gentzkow (2008)), a set of outcome-preferences separate from the preferences revealed in ordinary choice.

5 Existing Data on Implicit Preferences

A small number of papers come close to measuring implicit preferences in the way we define. We discuss them here and show how our formalization leads, in each case, to an improved test for implicit preference.

5.1 Snyder et al. (1979) on discrimination

Snyder, Kleck, Strenta, and Mentzer (1979) (SKSM) report an experiment which compares direct and indirect choices as a "general strategy for detecting motives that people wish to conceal." Their name for this general phenomenon is "attributional ambiguity," and their description, although not formalized, comes very close to our basic analysis of revealingness and implicit preferences. Subjects were invited to choose between sitting in one of two booths, in each of which a movie was being shown. Each booth already contained another person, who was either seated or was sitting in a wheelchair. The treatments varied in whether the booths were labelled to show either the same, or different, movies. The paper found that when the movies were the same, 75% (18/24) of subjects sat with the handicapped confederate, while when they were different only 33% (8/24) chose to sit with the handicapped confederate, intuitively pointing to an implicit preference against sitting with the handicapped individual: they write "avoidance of the handicapped ... masquerade[d] as a movie preference."

However, this is not an appropriate test for implicit preferences, and in fact a rational decision-maker with strong preferences over movies and weak preferences over which confederate to sit with will exhibit the same pattern of choice. Instead we need to check for a figure-8 cycle, keeping in mind the appropriate triangle inequality (Regenwetter et al. (2011)) since the data are between-subjects. We find that the condition is not satisfied, i.e. the choices observed can be rationalized by conventional transitive preferences, and we provide an example. Collecting data on data on indifferences, collecting within-subjects data, or collecting data on evaluation may increase the ability to detect implicit preferences in this paradigm.

Subjects were invited to choose between sitting in one of two booths, in each of which a movie was being shown. Each booth already contained another person, who was either seated or was sitting in a wheelchair. The treatments varied in whether the booths were labelled to show either the same, or different, movies (the identities of the movies were cross-randomized against the assignment of the wheelchair). The

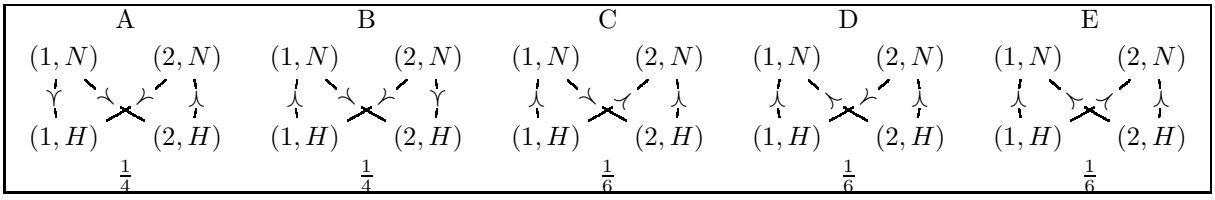
paper found that when the movies were the same, 75% (18/24) of subjects sat with the handicapped confederate, while when they were different only 33% (8/24) chose to sit with the handicapped confederate. SKSM test for a significant difference between the fractions, and conclude that "avoidance of the handicapped ... masquerade[d] as a movie preference."

However a simple difference in choice proportions does not establish that these preferences are inconsistent: a difference in proportions can occur if subjects have sufficiently strong preferences over which movie to see. Denoting movies by $i \in \{1, 2\}$ and confederates by $j \in \{N, H\}$ (Non-handicapped, Handicapped), SKSM test the null hypothesis that $Pr((1, H) \succ (2, N)) + Pr((2, H) \succ (1, N)) = Pr((1, H) \succ (1, N)) + Pr((2, H) \succ (2, N))$. Even when subjects have context-independent, transitive preferences this condition will be violated in all but knife-edge cases. For example, the condition is violated (a strict inequality $>$) if all subjects have the preferences $(1, N) \succ (1, H) \succ (2, N) \succ (2, H)$, i.e. they prefer to see a given movie with a non-handicapped person, but strongly prefer movie 1 to movie 2.

Instead, we need to check whether people exhibit a figure-8 cycle. As discussed earlier, for aggregate data on a 4-element choice cycle to be irreconcilable with heterogeneous, transitive preferences the average rate of preference must exceed $\frac{3}{4}$ (Regenwetter et al. (2011)). SKSM find only an average rate of preference along their cycle of 71% $((75\% + (1 - 33\%))/2)$. In other words, the data that they observe could be generated by a mixture of agents each of who has a transitive choice function.[53]

Collecting data on data on indifferences, collecting within-subjects data, or collecting data on evaluation may increase the ability to detect implicit preferences.

[53]For example the following 5 types would generate the observed choice fractions (denoting movies by $i \in \{1, 2\}$ and confederates by $j \in \{N, H\}$ (Non-handicapped, Handicapped)):



This hypothetical distribution of preferences would imply that 75% of subjects choose to sit with the handicapped person when the movies are the same (B, C, D, E for movie 1, and A, C, D, E for movie 2), and 33% sit with the handicapped person when the movies are different (C and E when the handicapped person is viewing movie 1, and D and E when they are viewing movie 2). If we assume that preferences are separable in the movie type, then transitivity of underlying preferences requires that the average rate of preference along the cycle exceed $\frac{2}{3}$, meaning the observed data does violate transitivity, although the difference is unlikely to be statistically significant. This discussion underlines the care that needs to be taken in testing for implicit preferences.

5.2 Exley (2015) on self-serving biases

Exley (2015) experimentally studies "excuse-driven risk preferences," and finds that risk-preferences seem to *change* in a self-serving way. Subjects choose between payoffs to charity and payoffs to themselves: when the payoff to charity has some risk, then decision-makers are risk averse; but when the payoff to themselves has risk, then decision-makers become relatively risk-loving. Her experimental design is the closest that we are aware of to the approach we propose in this paper, and we show that her data do indeed reveal implicit preferences: under a mild assumption her data reveal "two triangles" that identify an implicit preference for self-payoffs over charity-payoffs. Some subjects also exhibit a "figure-8" cycle that reveals an additional implicit preference over risk: some subjects are implicitly risk-averse. Exley shows that excuse-driven behavior correlates with selfishness as well as the propensity to "wiggle" in a moral wiggle-room task (Dana et al. (2007)), suggesting that all three behaviors capture a common feature of preferences. This correlation is consistent with the intuition that "wiggle room" behavior can also be thought of under the banner of implicit preferences.

Subjects make five types of choices. In Exley's notation, she first elicits (using a choice list), the X that makes each individual subject indifferent between $(10, 0)$ and $(0, X)$, where $(10, 0)$ denotes \$10 for self, and \$0 for charity. Thereafter, she elicits four types of certainty equivalent. $Y^S(P^S)$ is the certain amount for self, expressed as a percentage of \$10, that makes the subject indifferent to a self-lottery P^S that pays \$10 to self with probability P^S, nothing otherwise. $Y^C(P^C)$ is the certain amount for charity, as a percentage of $\$X$, that yields indifference to a charity-lottery paying $\$X$ to charity with probability P^C. $Y^S(P^C)$ is the analogous self-dollar valuation of the charity-lottery, and $Y^C(P^S)$ the charity-dollar valuation of the self-lottery. Hence a higher value for $Y^i(P^j)$ implies a higher value assigned to that j-lottery when measured in i-dollars.

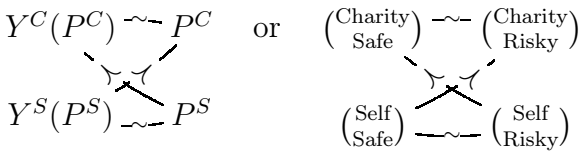
The basic choice behavior of interest, presented both graphically and in regressions, can be summed up by Exley's Prediction 3 (Excuse-driven preferences). It can be written as follows:

$$Y^C(P^S) > Y^S(P^S) = Y^C(P^C) > Y^S(P^C)$$

In words, a self-lottery is assigned a higher valuation in charity-dollars than in self-dollars, while a charity-lottery is assigned a lower valuation in self-dollars than in charity dollars. Yet the self-dollar valuation of the self-lottery is equal to the charity-

dollar valuation of the charity-lottery. This does appear intuitively consistent with an implicit preference for self, manifesting as context-specific risk preferences. Exley regresses $Y^i(P^j)$ on dummies for $\mathcal{I}[j = C]$ ("charity"), $\mathcal{I}[i \neq j]$ ("tradeoff"), and $\mathcal{I}[i = Charity] * \mathcal{I}[i \neq j]$ ("charity*tradeoff"), which allows her to check whether the average valuations satisfy the three inequalities in Prediction 3.

Perhaps surprisingly, the *observed choices alone* as described are consistent with transitivity. In order to establish intransitive cycles one must make an additional assumption, although one we think is reasonable. We translate the observed choices of an individual who is consistent with Prediction 3 into our graphical representation below:[54]



These choices represented are not intransitive, in our terminology they are consistent with a purely explicit preference for self, choosing the self-favoring option when available. Adding an assumption about preferences in the vertical choice sets allows us to establish two triangles. For example, adding $Y^S(P^S) \precsim Y^C(P^C)$ or $P^S \precsim P^C$ would be sufficient to establish a positive implicit preference for self. While subjects did face a choice between $(10, 0)$ and $(0, X)$, they did not face a choice between $(10 * Y^S(P^S), 0)$ and $(0, X * Y^C(P^C))$, and $Y^S(P^S) = Y^C(P^C)$ does not necessarily imply $Y^S(P^S) \sim Y^C(P^C)$ (e.g. it might be that $(9, 0) \succ (0, 0.9X)$). To establish inconsistency, Exley assumes that charity-dollar amounts can be translated into self-dollar amounts at exchange rate $10/X$, implying that $Y^C(P^C) = Y^S(P^S) \Leftrightarrow Y^C(P^C) \sim Y^S(P^S)$ and thus yielding two triangles that together reveal an implicit preference for self (triangle 1 is $Y^C(P^C) \sim P^C \succ Y^C(P^S) \sim Y^C(P^C)$ and triangle 2 is $Y^S(P^S) \sim P^S \succ Y^C(P^C) \sim Y^S(P^S)$).

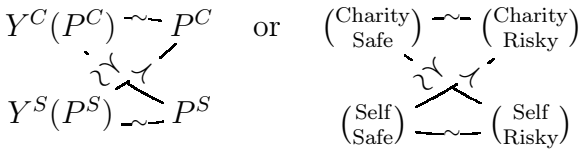
We do believe this is a reasonable assumption, because a) Exley detects strong effects even for probabilities P^i close to one, where it is reasonable to assume that approximate indifference between self and charity safe payoffs is preserved, and b) as Exley points out, by construction $P^C \sim P^S$ is implied by any theory of risk preferences that obeys the Independence axiom.[55] We raise the point because it determines what

[54]The horizontal indifferences follow from the fact that the horizontals are directly elicited via the choice list. The diagonals follow from the fact that, for example, $Y^C(P^S) > Y^C(P^C)$, and therefore we know by the structure of the choice list that the subject also chose P^S over $Y^C(P^C)$ when presented with the choice (because preferences are monotone in self-dollars and charity-dollars).

[55]Additionally, valuations are very close to risk neutral when valuing self-lotteries in self-dollars and charity-lotteries in charity-dollars, and furthermore when the linear translation is applied she finds that $Y^S(P^S) \approx Y^C(P^C)$ for all probabilities, i.e. linearity seems to do a good job of predicting behavior.

is, under our framework, the optimal approach to detecting implicit preferences in Exley's data. *With* the assumption, one need not run any regression. By exploiting the within-subjects nature of the data one can directly identify the individuals who exhibit the relevant cycles.[56]Meanwhile, pooled regressions (which compare average valuations) do not exploit the power of within-subject data and risk missing the presence of implicit preferences amidst the sampling variation. They also require stronger restrictions on the underlying heterogeneity of implicit preferences. *Without* the assumption either analysis risks mistaking transitive underlying preferences for implicit preferences. We note that an advantage of the pooled regressions is the ability to correlate the magnitude of deviations from risk-neutrality in the diagonal choice sets, with the tendency to exhibit "wiggle-room" behavior, and with selfishness (high value of X). A similar exercise would also be possible from a binary classification of subjects by their implicit preferences as identified in the within-subjects data.

Finally, we do also find some data on subject-level intransitive choices, reported in Exley's footnote 29: for $P = 0.95$, 42 percent of subjects made intransitive choices summarized by $Y^S(P^S) > Y^S(P^C)$ and $Y^C(P^S) \leq Y^C(P^C)$. Using $P^S \sim Y^S(P^S)$, $Y^C(P^C) \sim P^C$, and the structure of the choice set (which tells us, for example, that the subject chose $Y^S(P^S)$ over P^C when given the choice) we can translate the choices into our representation, revealing an intransitive figure-8 cycle:



These choices identify an implicit negative preference for *risk*: 42% of subjects are systematically more risk averse in the less revealing (diagonal) choice sets. An explanation could be that riskiness makes people uneasy, but they find it difficult to rationalize that uneasiness, unless the riskiness is bundled some other other attribute, as it is in the diagonal choice sets.

Taking the evidence as a whole, it seems that Exley's subjects do exhibit implicit preferences over Self-Charity, and we also find evidence for implicit preferences over Risky-Safe.

[56]Although Exley does not explicitly report how many individuals are cyclical, she tells us that 78% of choices exhibit $Y^i(P^j) \neq Y^j(P^j)$. Assuming that all of these differences go in the predicted direction, this implies that at least 78% of subjects revealed at least one of the diagonal preferences (at least one triangle), and that at least 56% of people exhibited both diagonals at least once for a given P (two triangles).

5.3 DeSante (2013) on racial discrimination

DeSante (2013) finds racial bias in an experiment where subjects are asked to set welfare payments for applicants who vary in various attributes. In his experiment two applicants are evaluated at once, allowing us to test for implicit preferences. Reanalyzing the data we find evidence for a negative implicit preference for black candidates, and additionally a negative implicit preference for candidates with high "work ethic".

In DeSante (2013) 1,000 subjects were asked to provide recommendations for state welfare payments to two hypothetical recipients - i.e. evaluation data on pairs of outcomes. Subjects were told they had \$1,500 which they should allocate between two applicants presented side-by-side, with any remainder to be "added to state funds." Applicants differ in race (Black, White) signaled by name, and "work ethic" (Poor, Excellent).[57] The design is not ideal because the budget constraint creates a trade-off between the two applicants, whereas true joint evaluation does not impose choice-set dependence. Nevertheless, it is the closest example to our proposed method of which we are aware. Since the paper does not perform our specific comparisons of interest, we obtained and reanalyzed the experimental data.

We focus on the half of the subjects who were shown applicants with both attributes.[58] For simplicity we ignore "background" attributes which are held constant, which is equivalent to assuming there are no implicit preferences over these attributes. There are four types of evaluation sets: $\{WE, BP\}$, $\{WE, WP\}$, $\{BE, WP\}$ and $\{BE, BP\}$. There exist two parallel scissors:

$$(1) \quad y(WE|BP) - y(WE|WP) \text{ and } y(BE|WP) - y(BE|BP)$$
$$(2) \quad y(WP|BE) - y(WP|WE) \text{ and } y(BP|WE) - y(BP|BE)$$

For (1) we find $y(WE|BP) < y(WE|WP)$, inconsistent with a weakly positive

[57] Latoya and Keisha for black applicants, Laurie and Emily for whites. Because only two names are used for each race, and in fact only one name for each race appears in the mixed-race evaluation sets it is difficult to conclusively separate implicit preferences over names from implicit preferences over race.

[58] The other half did not have work ethic information, so in our framework can be thought of as having only one attribute, race. This can be analyzed as a variant of joint/separate evaluation in our framework (the evaluation sets are $\{W, W\}, \{B, B\}, \{W, B\}$ where the "separate" sets contain two applicants of the same race) and actually yield stronger evidence consistent with an implicit bias (Blacks applicants are are allocated \$557 on average when alongside another black applicant, and \$600 on average alongside a white applicant, while white applicants receive \$583 when alongside a white and \$556 when alongside a black applicant), but we focus on the two-attribute treatment.

implicit preference for both Black and Excellent, and $y(BE|WP) < y(BE|BP)$, inconsistent with a weakly positive preference for both White and Excellent. Hence the data in (1) imply a positive implicit preference for Excellent, but are inconclusive about race.[59]

For (2) we find $y(WP|BE) < y(WP|WE)$, inconsistent with a weakly positive implicit preference for both Black and Poor and $y(BP|WE) > y(BP|BE)$, inconsistent with a weakly positive implicit preference for both Black and Excellent, and therefore implying a negative implicit preference for Black applicants.[60]

5.4 Bohnet et al. (2015) on gender preferences

A recent experimental paper by Bohnet et al. (2015) can be interpreted as studying implicit preferences. They study whether a decision-maker's choice between candidates for a task becomes more or less sensitive to certain attributes–gender and past performance–when the choice is either between an individual candidate and an unobserved "pool" alternative, or between two candidates and the pool (a paradigm closely related to joint and separate evaluation, although it elicits choices, not evaluations). They find that "disadvantaged gender" candidates are less likely to be selected when considered individually than when considered alongside an advantaged gender alternative. On the contrary, low ability candidates are more likely to be selected when considered individually than when the alternative is a high ability candidate.

While intuitively the variation in frequency of certain choices points to implicit preferences (as we have argued, considering multiple candidates increases revealingness with respect to their attributes), in fact we show that it is not possible to infer implicit preferences from these data: regular, transitive preferences will generate the patterns of choice that Bohnet et al. (2015)'s tests interpret as varying sensitivity. A simple

[59]Note that this test is *not* a valid parallel scissors in general for detecting an implicit preference over work ethic: the target in both scissors is Excellent, while parallel scissors would require one Poor and one Excellent target (see Proposition 5). We are nevertheless able to make a conclusive statement about work ethic preferences because we have only two attributes, and hence two scissors can be sufficient (see Appendix 8.3).

[60]For simplicity, as in our discussion of Snyder et al. (1979) we have treated the sample averages as though they were population averages. In fact, of the four scissors only $y(WE|BP) - y(WE|WP) < 0$ is statistically significant ($p = 0.052$). There are various approaches one could take for performing joint hypothesis tests, a non-trivial issue due to the compound directional hypotheses under consideration. For example, a standard F- or Chi-square test would test the joint null that all scissors are equal to zero, hence a rejection provides evidence for some choice-set dependence but not immediately about its form.

example illustrates the basic point: imagine a decision-maker whose preferences over quality are $High \succ Low \succ Pool$. Then they are "more sensitive" to quality in the choice set $\{High, Low, Pool\}$ (High is always chosen and Low is never chosen), than in $\{High, Pool\}$ and $\{Low, Pool\}$ (High and Low are equally likely to be chosen in each case). We do however show that Bohnet et al. (2015)'s subjects exhibit violations of WARP that point to implicit preferences, though we believe are harder to interpret. Instead, the most natural way to test for implicit preferences in their paradigm would be to collect *evaluation* data, and conduct our scissors tests.

In the experiment, lab subjects choose between candidates to perform a task, and are rewarded according to the chosen candidate's performance. Subjects observe the candidates' gender and prior performance (above/below average). They can also always choose instead to take a random draw from the pool of other candidates. Subjects face a single choice from one of two kinds of choice set: "separate" (a choice between one candidate or the pool), and "joint" (a choice between a male candidate and a female candidate, each with different level of prior performance, and the pool). Finally, Bohnet et al. (2015) also vary the type of task (math/verbal), to test whether gender bias is task-specific, with the maintained hypothesis that males are the "advantaged" gender in math, and females in verbal. To pool the data, they recode gender as "advantaged/disadvantaged." The paper is interested in whether subjects' choices reveal a gender influence that varies between treatments, with the hypothesis that the influence of gender decreases (and the influence of quality increases) in the joint treatment.

To aid comparison, we need to clarify terminology. Bohnet et al. (2015) refer to the two treatments as *separate evaluation* and *joint evaluation* respectively. In our terminology they are eliciting choices, not evaluations. While we will retain their use of "joint" and "separate," formally we are studying choices from either binary or ternary choice sets.

Bohnet et al. (2015)'s main analysis studies choice proportions, comparing the relative likelihood that the advantaged/disadvantaged candidate (averaging over quality) or high/low candidate (averaging over gender) is chosen, between joint and separate. They find that advantaged candidates are more likely to be chosen from "separate" choice sets, but approximately equally likely in joint, and they find the opposite pattern for high performance candidates. This is interpreted as showing that choices are more sensitive to gender, and less sensitive to quality, in the separate treatment. Next, they run a Probit regression estimating the likelihood that a candidate is chosen as a function of gender and quality. In the separate treatment the coefficient on gender

is large and significant and the coefficient on quality is small and not significant, and the reverse is found in the joint treatment. This is again interpreted as evidence for changing sensitivity between modes.

It is clear that the aggregate sensitivity of choice likelihood varies between joint and separate treatments, a fact that is surely of interest to policymakers. Interestingly, however, this analysis is *not* sufficient to conclude that subjects' underlying gender and quality preferences are choice-set dependent (or, specifically, implicit). To see this, observe that the test for changing sensitivity to quality has the following null (where gender is A/D, quality is H/L and Pool is "P"):

$$\begin{aligned}
&\frac{Pr(AH = c\{AH, DL, P\}) + Pr(DH = c\{DH, AL, P\})}{2} \\
&- \frac{Pr(AL = c\{AL, DH, P\}) + Pr(DL = c\{DL, AH, P\})}{2} \\
&= \frac{Pr(AH = c\{AH, P\}) + Pr(DH = c\{DH, P\})}{2} \\
&- \frac{Pr(AL = c\{AL, P\}) + Pr(DL = c\{DL, P\})}{2}
\end{aligned}$$

which simplifies to $Pr(AH \succ DL \succ P) + Pr(DH \succ AL \succ P) = Pr(DL \succ AH \succ P) + Pr(AL \succ DH \succ P)$. This condition will be violated–and hence a large sample test will reject with probability one–in all but knife-edge cases, a simple example would be where all subjects have $DH \succ AH \succ DL \succ AL \succ P$. Hence a rejection of the null is fully consistent with context-independent (transitive) underlying preferences.[61] An equivalent argument shows that the test for changing sensitivity to gender is also invalid.

One can construct populations of purely transitive individuals that replicate the pattern of Bohnet et al.'s results for both gender and quality. In Example 8 we show that conventional context-*independent* preferences can generate aggregate choice proportions which under the approach in Bohnet et al. would be interpreted as evidence for context-*dependent* preference.

[61] One can also argue that Bohnet et al.'s finding of higher sensitivity to quality in Joint (i.e. a strict inequality $>$) should be *expected*. Inspection of the condition reveals that for the opposite inequality ($<$) to hold, a relatively large fraction of subjects must exhibit strong preferences over gender such that a Low candidate of one gender is preferred to a High candidate of the other. Of course, knowing whether gender preferences are strong or weak is of interest, but is quite distinct from the hypothesis of context-dependent preferences.

Table 2: Heterogeneity and choice proportions

	Gender			Quality		
	A	D	Gap	H	L	Gap
Separate	75%	50%	25%	62.5%	62.5%	0%
Joint	37.5%	37.5%	0%	50%	25%	25%

Note: choice proportions constructed from example in the text.

Example 8. Suppose subjects had the following transitive preferences, in equal proportions: (1) $AH \succ AL \succ Pool \succ DH \succ DL$; (2) $DH \succ AH \succ DL \succ AL \succ Pool$; (3) $DH \succ DL \succ AH \succ AL \succ Pool$; (4) $Pool \succ DH \succ DL \succ AH \succ AL$. Averaging over the six possible choice sets in Bohnet et al. (2015)[62], we would observe an apparent gender advantage in separate that disappears in joint, and an apparent quality advantage in joint that disappears in separate (see Table 2). In other words, aggregate sensitivity to gender and quality varies, while the sensitivity of the underlying preferences does not. The assumed individual preferences actually tend to favor *disadvantaged* candidates: types 3 and 4 always rank D above A, while type 2 prefers D to A in 3 out of 4 pairwise comparisons.

In the regression analysis, the maintained assumption is that subjects' preferences over candidates can be characterized by a representative utility function, perturbed by independent (Gaussian or extreme-value distributed) shocks. It therefore cannot capture the correlation between preferences generated, for example, by the types in our above examples, and will tend to incorrectly identify instability in preferences if the underlying individual utilities are stable but correlated.

Turning to our own model, because we have defined it only over binary choice sets (with the exception of the discussion in 9), and binary attributes (which the "Pool" is not), it is difficult to prescribe the optimal test for implicit preferences Bohnet et al.'s data. However, there is one natural test for violations of rationality: one can check for violations of the Weak Axiom of Revealed Preferences (WARP). WARP requires that subjects do not become more likely to choose a given candidate when the choice set increases in size. Under the maintained assumption that deviations from rationality are generated only by implicit preferences, detecting such a violation confirms their existence (but does not, we believe, identify over which attribute). We find four WARP violations, of which one is marginally significant, detailed in Table 3.

[62]Specifically, $\{AH, Pool\}, \{DH, Pool\}, \{AL, Pool\}, \{DL, Pool\}, \{AH, DL, Pool\}, \{AL, DH, Pool\}$.

Table 3: WARP violations in Bohnet et al. (2015) data

	Joint choice	Separate choice	p-value
Math	$Pr(FH = c\{FH, ML, P\}) = 57\%$	$Pr(FH = c\{FH, P\}) = 44\%$	0.14
Math	$Pr(P = c\{FH, ML, P\}) = 40\%$	$Pr(P = c\{ML, P\}) = 35\%$	0.33
Math	$Pr(P = c\{MH, FL, P\}) = 42\%$	$Pr(P = c\{MH, P\}) = 34\%$	0.28
Verbal	$Pr(P = c\{FH, ML, P\}) = 38\%$	$Pr(P = c\{FH, P\}) = 19\%$	0.08*

Note: p-values from one-sided test of proportions.

6 Applications

In this section we discuss how certain anomalies in decision-making, across a variety of domains, can be interpreted as the expression of implicit preferences.

6.1 Implicit Discrimination

It has often been argued that preferences over race and sex are *implicit* in a way that other preferences are not: that they are hidden, or they are unconscious, or that they are immanent in the language that we use.

Since the mid 20th century it has become common, among philosophers and cultural theorists, to claim that our beliefs and preferences are subtly influenced by the culture we live in, in a way that is biased towards existing power structures. For example, that unspoken assumptions make it difficult to question existing class, sex, and race relations. Much intellectual work in Marxism, feminism, and race studies has tried to identify biases in different parts of everyday thought and culture. However the interpretation of the evidence, for example the analysis of texts, is disputable.

More recently an empirical case has been made for the implicitness of discrimination by comparing verbal reports of preference with actual behavior. This takes two forms: studies which find large differences in how people are treated, depending on their race or gender;[63] and studies which find differences in automatic associations.[64]

These approaches equate explicit preferences with stated preference, and implicit preference with revealed preference. Our claim is that we can identify *both* just from revealed preferences. Most closely related to our theory is Gaertner and Dovidio's

[63]See Mullainathan (2015) for a selection of studies which find large effects of race discrimination.

[64]Most famously the "Implicit Association Test," which finds that most people perform significantly better at a task which asks them to associate white faces with positive words, and black faces with negative words, than the opposite combination.

(1986) work on "aversive racism" - they argue that most people in the US are no longer overtly racist, but their judgment and decisions reflect racial influences in hidden ways.

Our theory has a simple implication for experimental design: by varying revealingness we can determine the degree to which discrimination is implicit. Existing designs can be extended by asking subjects to consider two outcomes instead of one - either simultaneously or in sequence. This can also be applied in field experiments, as long as it is reasonable to believe that the subject will find the two outcomes to be salient comparisons - for example, sending two CVs in application for a job, or sending two testers to apply for an apartment or mortgage.[65] Put simply: between-subject studies and within-subject studies are expected to show different outcomes, and the difference will tell us about implicit preferences.

If a large part of discrimination is implicit, in our sense, this implies that it will be more pronounced in situations that are less revealing. In particular, we would expect discrimination to be stronger when cases are evaluated one-by-one, than when they are evaluated in groups. Consider two hiring policies: one in which job applications are evaluated as they arrive, and one in which applications accumulate and are evaluated in groups. We expect differences in treatment to decline under the second policy.[66] There are also interesting implications of providing, to a decision-maker, aggregated information about their own decisions, for example providing a judge with data on the average prison term they have sentenced defendants of different races to. If the implicit discrimination is due to implicit knowledge, this information will help the decision-maker to learn about their own biases and adjust for them. If it is due to signaling, it could have the opposite effect because the marginal effect of a sentence on an observer's beliefs could decrease.[67] Finally, the theory characterizes the subjective experience of people who are discriminated against; as put by Snyder et al. (1979): "the handicapped person may be repeatedly rebuffed in social encounters by people who give what may seem to them to be reasonable excuses."

[65] We have piloted an experiment in which subjects are shown two defendants, and asked to suggest appropriate sentences, varying the race and crime used. Preliminary results find little explicit racial discrimination, and significant implicit racial discrimination.

[66] Our joint-separate result deals with groups of two. We discuss results for larger groups in the Appendix.

[67] This depends on the interpretation of the observer in the model - when judgments of n outcomes are aggregated, does the decision-maker care about the beliefs of n different observers?

6.2 Interpersonal Preferences.

Moral judgment is famously *opaque*: people find it easy to label actions as right or wrong, and fair or unfair, but hard to explain why they gave those labels. Much of moral philosophy proceeds by testing novel cases against intuition. These observations suggest that we have little direct introspection into our moral sense, and therefore that there could be large implicit effects. We make some suggestions of possible implicit influences, and discuss the relevant evidence that we are aware of.

Self-other tradeoffs. The most obvious implicit preference is a self-regarding bias: that people may put less weight on other peoples' payoffs, relative to their own, when the choice set becomes less revealing regarding that preference. This is a natural interpretation of the experiments in Exley (2015), who describes her results as "excuse driven." However we might also find the opposite implicit preference in some circumstances: Miller (1999) argues that contemporary American society exhibits a "norm of self-interest," which requires that people find a justification for their behavior on self-interested grounds: for example he claims that people are significantly more likely to contribute to charity when they are offered a trinket in exchange, because the exchange gives them a selfish excuse to perform a generous act.

Inequality aversion. A large literature has studied aversion to inequality inside and outside the lab. We believe that these preferences may be importantly implicit: i.e., inequality may have a bigger effect on choice in less revealing contexts. An indication of this is found in an experiment by Bazerman et al. (1992) which asked subjects to rate the fairness of two different allocations of money:

$$\binom{\text{self=\$500}}{\text{neighbour=\$500}} \tag{1}$$

$$\binom{\text{self=\$600}}{\text{neighbour=\$800}} \tag{2}$$

They found that when the outcomes were presented separately then the subjects rated (1) more highly than (2), but when they were presented jointly the ranking reversed. A loose interpretation of these results is that people dislike getting less than their neighbor (as occurs in 2), but that preference is implicit, and so its influence diminishes in joint evaluation.

Emotional/aesthetic aspects of a recipient. Patterns of giving to charity are famously difficult to reconcile with consequentialist preferences. We expect that peoples' implicit and explicit preferences regarding charity are quite different. As an illustration Kahneman and Ritov (1994) report that subjects rated a charity devoted to "skin cancer research" higher than one devoted to "saving Australian mammals." However when the charities were evaluated separately the average rating was higher for the latter (Kahneman and Ritov report a series of other similar reversals among charities).

Other influences. Schwitzgebel and Cushman (2012) report experimental results showing that judgments of moral responsibility are influenced by features which are often thought to be normatively irrelevant: whether the action is described as active or passive (action/omission); whether harm caused is a side-effect of aiming at a good outcome (the doctrine of double effect); and whether the outcome is under the decision-maker's control (moral luck). They additionally find that judgment is affected by the *order* of presentation: when asked about two situations, which vary only in one of these normatively-irrelevant features, respondents maintain consistency with their first answer. We therefore interpret their findings as establishing implicit preferences for these features.

6.3 Framing Effects

A framing effect is usually thought of as an influence on choice by a normatively irrelevant feature of the choice context (Tversky and Kahneman (1981)). Typical examples of framing effects are (1) the position of a reference point used in describing an outcome; (2) the position of an irrelevant anchor; (3) the designation of which alternative is the 'default' alternative; and (4) whether different aspects of an outcome are described separately or combined. However in each of these cases it is arguable whether the feature is indeed normatively irrelevant - the decision-maker may have preferences over that feature, or consider the feature informative.

An alternative definition - which does not require an assumption about which features are normatively relevant - can be given using our framework: a frame is an attribute over which there is an implicit preference, but no explicit preference. Any framing effect can therefore be described with an intransitive cycle. Some typical framing effects are represented in the following isosceles cycles.[68]

[68]The effect of gamble frame is discussed in Levin et al. (1987). The choices with cards are reported

$$
\begin{array}{ccccccc}
z & \succ & x & \precsim & x' & \succ & z \\
\$1 & \succ & \begin{pmatrix}\text{gamble}\\ \text{positive frame}\end{pmatrix} & \sim & \begin{pmatrix}\text{gamble}\\ \text{negative frame}\end{pmatrix} & \succ & \$1 \\
\$1 & \succ & \begin{pmatrix}\text{10 good cards}\\ \text{3 bad cards}\end{pmatrix} & \succ & (\text{10 good cards}) & \succ & \$1 \\
\$5 & \succ & \begin{pmatrix}\text{8oz ice-cream}\\ \text{in 9oz cup}\end{pmatrix} & \succ & \begin{pmatrix}\text{7oz ice-cream}\\ \text{in 5oz cup}\end{pmatrix} & \succ & \$5
\end{array}
$$

Our proposed definition does not fit all cases in the literature because sometimes a frame works at the level of the choice set, not at the level of an individual outcome. Consider the anchoring effect: it does not make much sense to ask a subject to separately state their WTP for two identical goods, one of which has been anchored at price p_1, another which has been anchored at price p_2 - here the anchor seems to affect the entire choice set, not an individual outcome.

6.4 Implicit Preferences & Consumer Behavior

Consumer choice often involves choosing among *bundles* of attributes, and therefore revealingness will vary across consumption contexts. The methods used in this paper could be applied to consumption data, for example determining whether features of a house (bedrooms, hot tub, ocean view, central heating) have different implicit and explicit values.

Suppose consumers implicitly desire some product, in the sense that they have a positive implicit but a negative explicit preference it. Then the firm selling it will wish to make the purchase less revealing by bundling their product with other choices, for example bundling pornography with journalism, to make the purchase less revealing. Suppose instead that consumers implicitly *dislike* a product. Then the firm will wish to make the purchase *more* revealing by removing excuses to not buy the product.

Under the implicit knowledge model firms will also wish to bundle their product with attributes that the consumer knows to be valueless, but which evoke positive associations. Insofar as consumers are imperfectly aware of those associations they will attribute some of the positive feelings evoked to the true quality of the product.[69]

in List (2002), the choices with ice creams are discussed in Hsee and Zhang (2010). Each could also be described in a binary space, though somewhat less naturally.

[69]This is elaborated on in Cunningham (2014).

7 Conclusion

Given data on choices economists have mapped out a landscape of tastes and aversions across the utility function. The classical utility function has been modified, in many areas, to accommodate observed choices: for example, tastes and aversions regarding ambiguity, losses and gains, inequality, and relative consumption.[70]

However we believe that there is a meaningful sense in which behavior can be inconsistent with *any* stable set of preferences - that people struggle with different motivations, and that the effects of these struggles can be detected in choice data. Of course any choices can be explained with a utility function if the space of outcomes is sliced thin enough. What we mean is that meaning that positing an invariant utility function is not the most parsimonious way of explaining observed choices. We think of this paper as a contribution towards formalizing, in a relatively nonparametric way, this results of this struggle.[71] We suspect that many preferences that are strong in direct comparisons will become weak in indirect comparisons - such as preferences over equality of payoffs, preferences over ambiguity (Fox and Tversky, 1977), and preferences over small risks. We suspect that many preferences that are weak in direct comparison will become strong in indirect comparisons - such as preferences over race and sex, preference for relative status, and partisan political preferences.

We interpret some prior work as identifying this same struggle through different means: through injecting randomness into choice sets; through identifying intransitive choices; and through the many manipulations of choice environment - with cognitive load, time limits, or affective stimulation.

The basic intuition underlying our paper - that implicit attitudes are revealed in indirect comparisons - has appeared in prior literature. However our discussion of existing work shows how difficult it can be to properly identify these effects, and we argue that our framework can serve as basis for systematic investigations of implicit preferences.

[70] The same argument can be made substituting 'belief' for 'preference'.

[71] We think of Rubinstein (1988), Hsee (1996), and Cherepanov et al. (2013) as contributions to the same line of thought.

References

Ainslie, G. (1992). *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. New York: Cambridge University Press.

Andreoni, J., J. M. Rao, and H. Trachtman (2011, December). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. Working Paper 17648, National Bureau of Economic Research.

Bazerman, M. H., G. F. Loewenstein, and S. B. White (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 220–240.

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

Bénabou, R. and J. Tirole (2004). Willpower and personal rules. *Journal of Political Economy 112*(4), 848–886.

Bénabou, 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.

Bodner, R. and D. Prelec (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions 1*, 105–26.

Bohnet, I., M. H. Bazerman, and A. Van Geen (2015). When performance trumps gender bias: Joint versus separate evaluation. *Management Science, forthcoming*.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience and consumer choice. Technical report, National Bureau of Economic Research.

Brocas, I. and J. Carrillo (2008). The brain as a hierarchical organization. *The American Economic Review 98*(4), 1312–1346.

Bushong, B., M. Rabin, and J. Schwartzstein (2014). A model of relative thinking.

Busse, M. R., D. G. Pope, J. C. Pope, and J. Silva-Risso (2013). The overinfluence of weather fluctuations on convertible and 4-wheel drive purchases. University of Chicago Working Paper.

Caplin, A. and D. Martin (2011). A testable theory of imperfect perception. Working paper 17163, National Bureau of Economic Research.

Chance, Z. and M. I. Norton (2009). I read playboy for the articles. *The Interplay of Truth and Deception: New Agendas in Theory and Research*, 136.

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

Cunningham (2014). Biases and implicit preferences. Technical report, Institute for International Economic Studies.

Cunningham, T. (2012). Comparisons and choice. *Working Paper*.

Cunningham, T. E. (2013). Biases and implicit knowledge. Working Paper.

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.

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

Ellenberger, H. F. (1970). The discovery ofthe unconscious. *New York, Basic Books*.

Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies, forthcoming*.

Gelman, A. (2014).

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.

Greenwald, A. G., T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji (2009). Understanding and using the implicit association test: Iii. meta-analysis of predictive validity. *Journal of personality and social psychology 97*(1), 17.

Hanson, R. (2008). Hanson on signaling. *EconTalk*.

Hirshleifer, D. (2001). Investor psychology and asset pricing. *The Journal of Finance 56*(4), 1533–1597.

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. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes 67*(3), 247–257.

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

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.

John, L. K., G. Loewenstein, A. Acquisti, and J. Vosgerau (2013). Paradoxical effects of randomized response techniques.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

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

Kahneman, D. and I. Ritov (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty 9*(1), 5–37.

Kőszegi, B. and A. Szeidl (2011). A model of focusing in economic choice. Working Paper.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics 4*(1), 136–63.

Levin, I., R. Johnson, and M. Davis (1987). How information frame influences risky decisions: Between-subjects and within-subject comparisons* 1. *Journal of economic psychology 8*(1), 43–54.

List, J. (2002). Preference reversals of a different kind: The 'more is less' phenomenon. *The American Economic Review 92*(5), 1636–1643.

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

Mazar, N., B. Koszegi, and D. Ariely (2013). True context dependent preferences? the causes of market dependent valuations. *Journal of Behavioral Decision Making 27*(3), 200–208.

Mijović-Prelec, D. and D. Prelec (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1538), 227–240.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Ph. D. thesis, George Washington University.

Mullainathan, S. (2015, January). Racial bias, even when we have good intentions. *The New York Times*.

Newell, B. R. and D. R. Shanks (forthcoming). Unconscious influences on decision making: a critical review. *Behavioral and Brain Sciences*.

Nisbett, R. E. and T. D. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review 84*(3), 231–259.

Nosek, B. A., C. B. Hawkins, and R. S. Frazier (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences 15*(4), 152 – 159.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases. *Department of Economics, UCB*.

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

Rubinstein, A. (1988). Similarity and decision-making under risk (is there a utility theory resolution to the allais paradox?). *Journal of Economic Theory 46*(1), 145–153.

Simonsohn, U. (2010). Weather to go to college. *The Economic Journal 120*(543), 270–280.

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.

Spence, M. (1973). Job market signaling. *The quarterly journal of Economics*, 355–374.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458.

Veblen, T. (1899). *Theory of the Leisure Class*. Norwalk: Easton.

Von Hippel, W. and R. Trivers (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences 34*(01), 1–16.

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