Identifying Prediction Mistakes in Observational Data *

Ashesh Rambachan†

August 26, 2021
Preliminary. Comments welcome.

Abstract

Decision makers, such as judges, doctors, and managers, make consequential choices based on predictions of unknown outcomes. Do these decision makers make systematic prediction mistakes based on the available information? In empirical settings, the preferences and information sets of decision makers are unknown to researchers, which makes uncovering systematic prediction mistakes a difficult identification problem. I develop an econometric framework to tackle this challenge and provide conditions under which systematic prediction mistakes can be identified. I show that exclusion restrictions on which observable characteristics of decisions may directly affect the decision maker's preferences and quasi-experimental variation together are sufficient to identify systematic prediction mistakes. Based on these identification results, I develop a tractable test for whether a decision maker makes systematic prediction mistakes, and methods to characterize the types of systematic prediction mistakes being made. These results are applicable to empirical settings such as pretrial release, medical testing, and many others.

*First version: June 2021. I am especially grateful to Isaiah Andrews, Sendhil Mullainathan, Neil Shephard, Elie Tamer and Jens Ludwig for their invaluable feedback, support, and advice. I also thank Alex Albright, Nano Barahona, Laura Blattner, Iavor Bojinov, Raj Chetty, Will Dobbie, Xavier Gabaix, Matthew Gentzkow, Ed Glaeser, Bryan Graham, Emma Harrington, Larry Katz, Ross Mattheis, Robert Minton, Ljubica Ristovska, Jonathan Roth, Joshua Schwartzstein, Jesse Shapiro, and participants at the Brookings Institution's Artificial Intelligence Author's Conference for many useful comments and suggestions. I gratefully acknowledge financial support from the NSF Graduate Research Fellowship (Grant DGE1745303). All errors are my own.

†Harvard University, Department of Economics: asheshr@g.harvard.edu

1 Introduction

Decision makers, such as judges, doctors, and managers, are commonly tasked with making consequential choices based on a prediction of an unknown outcome. For example, in deciding whether to detain a defendant awaiting trial, a judge predicts what the defendant will do if released based on detailed information such as the defendant's current criminal charge and prior criminal record. Are these decision makers making systematic prediction mistakes based on the available information? This foundational question in behavioral economics and psychology (e.g., Meehl, 1954; Tversky and Kahneman, 1974) has renewed policy relevance and empirical life as machine learning based models increasingly replace or inform decision makers in criminal justice, health care, labor markets, and consumer finance.[1]

In assessing whether such data-driven models improve outcomes, empirical researchers in economics and computer science intuitively evaluate decision makers' implicit predictions through ad hoc comparisons of their choices against those made by machine learning based models (e.g., Kleinberg et al., 2015, 2018).[2] While such comparisons of decision makers against machine learning based models are appealing, uncovering whether a decision maker is making systematic prediction mistakes is challenging as both the decision maker's preferences and information set are unknown to us. Perhaps the decision maker's choices diverge from the model simply because she has preferences that differ from the model's objective function or observes additional information that is unavailable to the model. For example, we do not know how judges assess the cost of pretrial detention and judges may uncover useful information through their courtroom interactions with defendants but we do not observe these interactions. It is therefore essential to understand whether systematic prediction mistakes are identifiable in the face of this challenge under as weak as possible assumptions on the decision maker's preferences, information set, and resulting choice behavior.

In this paper, I tackle this challenge by developing an econometric framework to test whether

[1] Risk assessment tools are used in criminal justice systems throughout the United States (e.g., Stevenson, 2018; Albright, 2019; Dobbie and Yang, 2019; Stevenson and Doleac, 2019; Yang and Dobbie, 2020). Clinical risk assessments aid doctors in their testing and treatment decisions (e.g., Beaulieu-Jones et al., 2019; Abaluck et al., 2020; Chen et al., 2020*a*,*b*). For applications in consumer finance, see Einav, Jenkins and Levin (2013); Fuster et al. (2018); Gillis (2019); Dobbie et al. (2020); Blattner and Nelson (2021) in economics and Khandani, Kim and Lo (2010); Hardt, Price and Srebro (2016); Liu et al. (2018); Coston, Rambachan and Chouldechova (2021) in computer science. For discussions of workforce analytics and resume screening software, see Autor and Scarborough (2008); Jacob and Lefgren (2008); Feldman et al. (2015); Hoffman, Kahn and Li (2018); Erel et al. (2019); Raghavan et al. (2020); Frankel (2021).

[2] See, for example, Berk, Sorenson and Barnes (2016); Chalfin et al. (2016); Chouldechova et al. (2018); Cowgill (2018); Hoffman, Kahn and Li (2018); Erel et al. (2019); Ribers and Ullrich (2019); Li, Raymond and Bergman (2020); Jung et al. (2020*a*); Mullainathan and Obermeyer (2020); Jansen, Nguyen and Shams (2021). Comparing a decision maker's choices against a statistical model has a long tradition in psychology (e.g., Dawes, 1971, 1979; Dawes, Faust and Meehl, 1989; Camerer and Johnson, 1997; Grove et al., 2000; Kuncel et al., 2013). See Camerer (2019) for a modern review of this literature.

a decision maker makes systematic prediction mistakes, and if so characterize properties of the systematic prediction mistakes that are being made. I consider empirical settings, such as pretrial release, medical testing, hiring, and many others, in which a decision maker must make decisions for many individuals based on a prediction of some unknown outcome using each individual's characteristics. The available data on the decision maker's choices and associated outcomes suffer from a missing data problem (Heckman, 1974; Rubin, 1976; Heckman, 1979; Manski, 1989): we only observe the outcome conditional on the decision maker's choices (e.g., we only observe a defendant's behavior upon release if a judge released them). I explore the restrictions imposed on the decision maker's choices by expected utility maximization behavior, where the decision maker maximizes some (unknown to the researcher) utility function at beliefs about the outcome given the observable characteristics as well as some private information. The decision maker's beliefs given the observable characteristics are required to lie in the researcher's identified set for the conditional distribution of the outcome given the observable characteristics (what I call "accurate beliefs").[3] For this reason, if there exists no such preferences and information set that rationalize the decision maker's choices, I say the decision maker is making systematic prediction mistakes based on the observable characteristics.

I derive a sharp characterization, based on revealed preference inequalities over the available data, of the identified set of utility functions at which the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs. If these revealed preference inequalities are satisfied at a candidate utility function, then some distribution of private information can be constructed such that the decision maker cannot do strictly better than her observed choices in an expected utility sense. If the identified set of utility functions is empty, then the decision maker is making systematic prediction mistakes as there is no combination of preferences and private information at which her observed choices are consistent with expected utility maximization at accurate beliefs.

I prove that without further assumptions systematic prediction mistakes are *untestable* in the leading empirical case of a binary choice and binary outcome. If either all observable characteristics of individuals directly affect the decision maker's utility function or the missing data can take any value, then the identified set of utility functions is non-empty. Any observed variation in the decision maker's choices can be rationalized under accurate beliefs by a utility function and private information that richly vary across all the observable characteristics. However, placing an exclusion restriction on which observable characteristics may directly affect the decision maker's utility and leveraging a research design to construct informative bounds on the missing data re-

[3] Since the conditional distribution of the outcome given the observable characteristics may not be point identified due to the missing data problem, this is the sharpest restriction on the decision maker's beliefs that can be placed by the researcher.

stores the testability of expected utility maximization behavior at accurate beliefs. In this case, there are testable restrictions on the variation in the decision maker's choices across observable characteristics that do not directly affect the decision maker's utility. These testable restrictions ask: do there exist values of the remaining excluded characteristics at which the decision maker could exchange her choices and do strictly better at all values of the missing data satisfying the researcher's constructed bounds? Identifying systematic prediction mistakes therefore requires *both* behavioral assumptions about the decision maker's preferences *and* econometric methods to address the missing data problem.

I then show that testing these restrictions is equivalent to a moment inequality problem, which is a well-studied testing problem in econometrics and many procedures are available (e.g., Canay and Shaikh, 2017; Molinari, 2020). The number of moment inequalities grows with the dimensionality of the observable characteristics, which will typically be quite large in empirical applications. To deal with this practical challenge, I discuss how supervised machine learning methods may be used to perform valid dimension reduction on this testing problem. Researchers may construct a prediction function for the outcome on held out data and partition the observable characteristics into percentiles of predicted risk based on this estimated prediction function. Testing these implied revealed preference inequalities across percentiles of predicted risk is a valid test of the joint null hypothesis that the decision maker's choices maximize expected utility at preferences that satisfy the conjectured exclusion restriction and accurate beliefs. These results provide, to my knowledge, the first microfounded procedure for using such statistical models to identify systematic prediction mistakes in empirical settings that suffer from a missing data problem.

With this framework in place, I further establish that the available data are also informative about the types of systematic prediction mistakes that are made by the decision maker. I extend the behavioral model to allow the decision maker to have possibly inaccurate beliefs about the unknown outcome (i.e., no longer requiring that the decision maker's beliefs lie in the identified set for the conditional distribution of the outcome given the observable characteristics). I sharply characterize the identified set of utility functions at which the decision maker's choices are consistent with "inaccurate" expected utility maximization and derive provide bounds on an interpretable parameter that summarizes the extent to which the decision maker's beliefs overreact or underreact to the observable characteristics. For a fixed pair of observable characteristics, these bounds summarize whether the decision maker's beliefs about the outcome vary more ("overreact") or less than ("underreact") the true conditional distribution of the outcome across these values. These bounds arise because any variation in the decision maker's choice probabilities across characteristics that do not directly affect utility must only arise due to variation in the decision maker's beliefs. Therefore, comparing variation in the decision maker's choice probabilities against variation in the probability of the outcome given the characteristics (which is partially identified) is informative

about the extent to which the decision maker's beliefs are inaccurate.

Taken together, these results provide a tractable econometric framework, which is rooted in economic theory, to study systematic prediction mistakes in a wide range of real-world empirical settings, such as pretrial release, medical testing, hiring, and many others. It both makes precise the behavioral assumptions and econometric assumptions that are required to identify systematic prediction mistakes, and characterizes the testable implications of systematic prediction mistakes. Furthermore, these results illustrate that real-world empirical settings can serve as rich laboratories for behavioral economics, enabling researchers to explore the behavior of human decision makers in high stakes domains. In future drafts, I will apply this econometric framework to analyze pretrial release decisions in the criminal justice system.

Related Literature: A large empirical literature evaluates decision makers' implicit predictions through either comparisons of their choices against those made by machine learning based models or estimating structural models of decision making. While the challenges of unknown preferences and information sets are recognized, researchers typically resort to strong assumptions that are tailored to particular empirical settings. Kleinberg et al. (2018) and Mullainathan and Obermeyer (2020) restrict preferences to be constant across both decisions and decision makers, and study average decision making behavior across a group of decision makers. Lakkaraju and Rudin (2017), Chouldechova et al. (2018), and Jung et al. (2020*a*) assume that observed choices were as-good-as randomly assigned given the characteristics, eliminating the problem of unknown information sets. Recent work introduces parametric models for the decision maker's private information, such as Abaluck et al. (2016), Arnold, Dobbie and Hull (2020), Chan, Gentzkow and Yu (2020), and Jung et al. (2020*b*). See also Currie and Macleod (2017), Ribers and Ullrich (2020), and Marquardt (2021). I develop an econometric framework for studying systematic prediction mistakes that only requires exclusion restrictions on which characteristics directly affect the preferences of a decision maker but otherwise places no further restrictions on preferences. It models the decision maker's information environment fully nonparametrically. This framework enables researchers to answer more questions under weaker assumptions, such as: Which decision makers make systematic prediction mistakes? What types of systematic prediction mistakes are being made by a decision maker? How do prediction mistakes vary across decision makers?

My identification results build on a literature that derives the testable implications of various behavioral models in "state-dependent stochastic choice (SDSC) data" (Caplin and Martin, 2015; Caplin and Dean, 2015; Caplin, 2016; Caplin et al., 2020). SDSC data contain full information on the joint distribution of a decision maker's choices and resulting outcomes, and therefore do not suffer from a missing data problem. While useful in analyzing lab-based experiments, such characterization results have limited applicability due to the difficulty of collecting such SDSC

data in practice (Gabaix, 2019; Rehbeck, 2020). I focus on common empirical settings in which the data suffer from a missing data problem, and I show that these settings can approximate ideal SDSC data by using quasi-experimental variation to address the missing data problem.

My analysis follows in the spirit of a literature on robust information design (e.g., Kamenica and Gentzkow, 2011; Bergemannn and Morris, 2013, 2016, 2019; Kamenica, 2019) by asking whether there exists *any* private information such that the decision maker's choices are consistent with expected utility maximization behavior. Syrgkanis, Tamer and Ziani (2018) and Bergemann, Brooks and Morris (2019) use results from this literature to study auctions and other multi-player games, whereas I analyze the choices of a single decision maker. Gualdani and Sinha (2020) also analyzes single-agent, discrete-choice settings under weak assumptions on the decision maker's information environment. Martin and Marx (2021) study the identification of taste-based discrimination by a decision maker in a binary choice experiments, providing bounds on the decision maker's group-dependent threshold rule. The setting I consider nests theirs by allowing for several key features of observational data such as missing outcomes, multi-valued outcomes, and multiple choices.

This paper relates to a literature on stochastic choice and information in decision theory such as Gul and Pesendorfer (2006), Ahn and Sarver (2013), Lu (2016), and Natenzon (2019). This literature typically assumes that the researcher observes the decision maker's random choice rule, which summarizes her choice probabilities in each possible menu of actions. I consider empirical settings in which we only observe a decision maker's choice probabilities from a single menu. Lu (2019) shows that a decision maker's state-dependent utilities and beliefs can be separately identified provided that the researcher observes choice probabilities across multiple informational treatments.

This paper complements recent results on the testability Roy-style selection (Mourifie, Henry and Meango, 2019; Henry, Meango and Mourifie, 2020) and the validity of "marginal outcome tests" for discrimination (Bohren et al., 2020; Canay, Mogstad and Mountjoy, 2020; Gelbach, 2021; Hull, 2021). The expected utility maximization model can be interpreted as a generalized Roy model, and it reduces to a generalization of the extended Roy model analyzed in Canay, Mogstad and Mountjoy (2020) and Hull (2021) in the special case of a binary decision and binary outcome. I formalize these connections in Section 2.

Finally, analyzing whether decision makers make prediction mistakes also has a long tradition in financial economics and macroeconomics. In financial economics, researchers study whether investors' expectations about asset prices satisfy rational expectations. See Shiller (1981), De Bondt and Thaler (1985), Campbell and Shiller (1987), and more recently Augenblick and Lazarus (2020), D'Haultfoeuille, Gaillac and Maurel (2020). In macroeconomics, researchers investigate whether households and professional forecasters have rational expectations about macroeconomic

aggregates (e.g., Sims, 2003; Carroll, 2003; Elliott, Timmerman and Komunjer, 2005; Elliott, Komunjer and Timmerman, 2008; Woodford, 2013).

2 Expected Utility Maximization in Screening Decisions

A decision maker makes choices for many individuals based on a prediction of an unknown outcome using each individual's characteristics. Under what conditions do the decision maker's choices maximize expected utility at some (unknown to us) utility function and accurate beliefs given the characteristics and private information?

2.1 The Observable Data

A decision maker faces many *screening decisions*, in which she makes choices for many individuals based on predictions of an unknown outcome. Each individual is summarized by characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$ and an outcome $y^* \in \mathcal{Y}$. The characteristics and outcome have finite support with $d_w := |\mathcal{W}|, d_x := |\mathcal{X}|$. For each individual, the decision maker selects a choice $c \in \{0, 1\}$. The random vector $(W, X, C, Y^*) \sim P$ defined over $\mathcal{W} \times \mathcal{X} \times \{0, 1\} \times \mathcal{Y}$ summarizes the joint distribution of the characteristics, the decision maker's choices, and the outcome across all individuals. I assume $P(W = w, X = x) \geq \delta$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ for some $\delta > 0$ throughout the paper.

The researcher observes the characteristics of each individual as well as the decision maker's choice. There is a *missing data problem*: the outcome Y^* is only observed whenever the decision maker selected $C = 1$. Defining the observable outcome as $Y := C \cdot Y^*$, the researcher observes the joint distribution $(W, X, C, Y) \sim P$. I assume the researcher knows this population distribution with certainty in order to focus on the identification challenges in this setting. Therefore, the researcher observes the decision maker's *conditional choice probabilities*

$$P(C = c \mid W = w, X = x) \text{ for all } c \in \{0, 1\}, (w, x) \in \mathcal{W} \times \mathcal{X}$$

as well as the *conditional choice-dependent outcome probabilities* given that the decision maker chose $C = 1$

$$P(Y^* = y^* \mid C = 1, W = w, X = x) \text{ for all } y^* \in \mathcal{Y}, (w, x) \in \mathcal{W} \times \mathcal{X}.$$

Importantly, the conditional choice-dependent outcome probabilities given $C = 0$, $P(Y^* = y^* \mid C = 0, W = w, X = x)$, are not observed due to the missing data problem.[4]

Notation: For a finite set $\mathcal{A}$, let $\Delta(\mathcal{A})$ denote the set of all probability distributions on $\mathcal{A}$ throughout the paper. I adopt the shorthand $P_{Y^*}(y^* \mid c = 0, w, x) := P(Y^* = y^* \mid C = 0, W =$

[4] I adopt the convention that $P(Y^* = y^* \mid C = c, W = w, X = x) = 0$ if $P(C = c \mid W = w, X = x) = 0$.

$w, X = x)$, $P_{Y^*}(y^* \mid c = 1, w, x) := P(Y^* = y^* \mid C = 1, W = w, X = x)$ and write $P_{Y^*}(\cdot \mid c = 0, w, x), P_{Y^*}(\cdot \mid c = 1, w, x) \in \Delta(\mathcal{Y})$ to denote the vectors of choice-dependent outcome probabilities given $C = 0, C = 1$ and observable characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$. Analogously, I write $P_{Y^*}(y^* \mid w, x) := P(Y^* = y^* \mid W = w, X = x)$ and $P_{Y^*}(\cdot \mid w, x) \in \Delta(\mathcal{Y})$ to denote the distribution of the outcome given observable characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$.

Example: Pretrial Release A judge decides whether to detain defendants awaiting trial (Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018; Arnold, Dobbie and Hull, 2020). The outcome $Y^* \in \{0, 1\}$ is whether a defendant would commit pretrial misconduct if released and the judge's choices $C \in \{0, 1\}$ are to either detain or release a defendant. The characteristics (W, X) summarize various information about the defendant that is available at the pretrial release hearing such as demographic information (race, gender, age, etc.), the current charges filed against the defendant and the defendant's prior criminal record. The researcher observes the characteristics of each defendant, whether the judge released them, and whether the defendant committed pretrial misconduct if the judge released them. The judge's conditional release rate $P(C = 1 \mid W = w, X = x)$ and conditional pretrial misconduct rate among released defendants $P(Y^* = 1 \mid C = 1, W = w, X = x)$ are observed. The conditional pretrial misconduct rate among detained defendants $P(Y^* = 1 \mid C = 0, W = w, X = x)$ is unobserved. ▲

Example: Medical Testing A doctor decides whether to conduct a costly medical test on patients (Abaluck et al., 2016; Ribers and Ullrich, 2019; Chan, Gentzkow and Yu, 2020). For example, shortly after an emergency room visit, a doctor decides whether to conduct a stress test on patients to determine whether they had a heart attack (Mullainathan and Obermeyer, 2020). The outcome $Y^* \in \{0, 1\}$ is whether the patient had a heart attack and the doctor's choices $C \in \{0, 1\}$ are whether to conduct the stress test. The characteristics (W, X) summarize rich information that is available about the patient such as demographics, reported symptoms, and prior medical history. The researcher observes the characteristics of each patient, whether the doctor conducted a stress test and whether the patient had a heart attack if the doctor conducted a stress test. The researcher observes the doctor's conditional stress testing rate $P(C = 1 \mid W = w, X = x)$ and the conditional heart attack rate among stress tested patients $P(Y^* = 1 \mid C = 1, W = w, X = x)$. The conditional heart attack rate among untested patients $P(Y^* = 1 \mid C = 0, W = w, X = x)$ is unknown. ▲

Example: Hiring A hiring manager decides whether to hire job applicants (e.g., Autor and Scarborough, 2008; Chalfin et al., 2016; Hoffman, Kahn and Li, 2018; Frankel, 2021).[5] The outcome

[5]The setting also applies to job interview decisions (Cowgill, 2018; Li, Raymond and Bergman, 2020), where the choice $C \in \{0, 1\}$ is whether to interview an applicant and the outcome $Y^* \in \{0, 1\}$ is whether the applicant is ultimately hired by the firm.

$Y^* \in \mathcal{Y}$ is some measure of on-the-job productivity, which for example may length of tenure since turnover is costly. The characteristics (W, X) are various information about the applicant such as demographics, education level and prior work history. The researcher observes the characteristics of each applicant, whether the manager hired the applicant, and their length of tenure if hired. The researcher observes the manager's conditional hiring rate $P(C = 1 \mid W = w, X = x)$ and the conditional distribution of tenure lengths among hired applicants $P(Y^* = y^* \mid C = 1, W = w, X = x)$. The researcher cannot observe the counterfactual distribution of tenure lengths among rejected applicants $P(Y^* = y^* \mid C = 0, W = w, X = x)$. ▲

Remark 2.1 (Extensions)**.** I make several simplifying assumptions in the main text of the paper for exposition. First, I focus on the case in which the decision maker faces only two choices. Second, I focus on the case in which the decision maker's choice does not have a direct causal effect on the outcome. Third, I assume the characteristics have finite support. I extend my results to settings in which the decision maker faces a *treatment assignment problem* with multiple choices and in which the characteristics are continuously distributed in Supplement D. Nonetheless, the results in the main text cover the empirical settings discussed above.

2.1.1 Bounds on the Missing Data

To address the missing data problem, I assume the researcher introduces additional assumptions or uses a research design to bound the unobservable choice-dependent outcome probabilities.

Assumption 2.1 (Bounds on Missing Data)**.** For each $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists a known subset $\mathcal{B}_{c=0,w,x} \subseteq \Delta(\mathcal{Y})$ such that $P_{Y^*}(\cdot \mid c = 0, w, x) \in \mathcal{B}_{c=0,w,x}$.

Let $\mathcal{B}_{c=0}$ denote the collection of bounds $\mathcal{B}_{c=0,w,x}$ at all $(w, x) \in \mathcal{W} \times \mathcal{X}$. In some cases, it may be useful to analyze the decision maker's choices without placing any further assumptions, which corresponds to setting $\mathcal{B}_{c=0,w,x} = \Delta(\mathcal{Y})$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$. In other cases, there may be an available research design that leverages quasi-experimental variation to produce bounds on the unobservable choice-dependent outcome probabilities, such as an instrumental variable or a proxy outcome. I return to this point in Section 3, where I show that common research designs naturally lead to bounds of this form.

Under Assumption 2.1, various features of the joint distribution $(W, X, C, Y^*) \sim P$ are partially identified. For example, the sharp identified set of the marginal distribution of the outcome given characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$, denoted by $\mathcal{H}_P(P_{Y^*}(\cdot \mid w, x); \mathcal{B}_{c=0,w,x}) \subseteq \Delta(\mathcal{Y})$, equals the set of all $\widetilde{P}_{Y^*}(\cdot \mid w, x) \in \Delta(\mathcal{Y})$ such that there exists $\widetilde{P}_{Y^*}(\cdot \mid c = 0, w, x) \in \mathcal{B}_{c=0,w,x}$ satisfying for all $y^* \in \mathcal{Y}$

$$\widetilde{P}_{Y^*}(y^* \mid w, x) = P(C = 1, Y^* = y^* \mid W = w, X = x) + \widetilde{P}_{Y^*}(y^* \mid c = 0, w, x) P(C = 0 \mid W = w, X = x)$$

If the bounds $\mathcal{B}_{c=0,w,x}$ are a singleton for all $(w, x) \in \mathcal{W} \times \mathcal{X}$, then the full joint distribution $(W, X, C, Y^*) \sim P$ is known and the observable data are equivalent to state-dependent stochastic choice data that is collected in lab-based experiments (e.g., Caplin, Dean and Martin, 2011; Caplin and Martin, 2015; Caplin and Dean, 2015; Caplin and Martin, 2020; Caplin et al., 2020). At the other extreme, the observable data are equivalent to stochastic choice data if neither choice generates observable outcomes.

2.2 Expected Utility Maximization Behavior

I examine the restrictions imposed on the decision maker's choices by expected utility maximization behavior, which models the decision maker as maximizing some (unknown to us) utility function at accurate beliefs about the outcome given the observable characteristics and some private information.

I define the two main ingredients of the expected utility maximization model. A utility function summarizes the decision maker's preferences over choices and outcomes, which may vary based on some of the observable characteristics. The decision maker's private information is some additional random variable $V \in \mathcal{V}$ that is available to the decision maker and predictive of the outcome of interest.

Definition 1. A *utility function* $U\colon \{0, 1\} \times \mathcal{Y} \times \mathcal{W} \to \mathbb{R}$ specifies the payoff associated with each choice-outcome pair, where $U(c, y^*; w)$ is the payoff associated with choice c and outcome y^* at characteristics $w \in \mathcal{W}$. Let $\mathcal{U}$ denote the feasible set of utility functions specified by the researcher.

Definition 2. The decision maker's *private information* is a random variable $V \in \mathcal{V}$.

Under the model, the decision maker observes the characteristics (W, X) as well as her private information $V \in \mathcal{V}$ prior to selecting a choice. The private information summarizes all additional information that is available to the decision maker but unobserved by the researcher. I make no assumption about the support of the private information. In empirical settings such as pretrial release and medical testing, researchers often worry that the decision maker observes additional private information that is not recorded in the observable data. Consequently, the decision maker's choices may reflect variation in this private information. Since it is unobservable to the researcher, a central goal of my analysis is to explore restrictions on the decision maker's behavior under weak, nonparametric assumptions on private information.

Based on this information set, the decision maker forms beliefs about the unknown outcome and selects a choice to maximize expected utility. Therefore, the expected utility maximization behavior model is summarized by a joint distribution over the observable characteristics, private information, choices and the outcome. I denote this joint distribution under the expected utility

maximization model as $(W, X, C, V, Y^*) \sim Q$. The decision maker's observed choices are consistent with expected utility maximization behavior if and only if there exists some utility function $U \in \mathcal{U}$ and joint distribution $(W, X, V, C, Y^*) \sim Q$ that satisfies several interpretable conditions and matches the observable data $(W, X, C, Y) \sim P$.

Definition 3. The decision maker's choices are *consistent with expected utility maximization behavior* if there exists a utility function $U \in \mathcal{U}$ and joint distribution $(W, X, V, C, Y^*) \sim Q$ satisfying

i. Information Set: $C \perp\!\!\!\perp Y^* \mid W, X, V$ under Q.

ii. Expected Utility Maximization: For all $c \in \{0, 1\}$, $(w, x, v) \in \mathcal{W} \times \mathcal{X} \times \mathcal{V}$ such that $Q(C = c \mid W = w, X = x, V = v) > 0$,

$$\mathbb{E}_Q \left[U(c, Y^*; W) \mid W = w, X = x, V = v\right] \geq \mathbb{E}_Q \left[U(c', Y^*; W) \mid W = w, X = x, V = v\right]$$

for $c' \neq c$.

iii. Data Consistency: For all $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists $\widetilde{P}_{Y^*}(\cdot \mid c = 0, w, x) \in \mathcal{B}_{c=0,w,x}$ satisfying for all $y^* \in \mathcal{Y}^*$

$$Q(W = w, X = x, C = 1, Y^* = y^*) = P(W = w, X = x, C = 1, Y^* = y^*),$$
$$Q(W = w, X = x, C = 0, Y^* = y^*) = \widetilde{P}_{Y^*}(y^* \mid c = 0, w, x) P(W = w, X = x, C = 0).$$

Definition 4. The *identified set of utility functions*, denoted by $\mathcal{H}_P(U; \mathcal{B}_{c=0}) \subseteq \mathcal{U}$, is the set of utility functions $U \in \mathcal{U}$ such that there exists a joint distribution $(W, X, V, C, Y^*) \sim Q$ satisfying Definition 3.

The decision maker's choices are consistent with expected utility maximization behavior if three conditions are satisfied. First, the decision maker's choices must be independent of the outcome given the characteristics and private information under the model Q ("Information Set"), formalizing the sense in which the decision maker's information set consists of only (W, X, V). Second, if the decision maker selects a choice c with positive probability given $W = w, X = x, V = v$ under the model Q, then it must have been optimal for her to do so ("Expected Utility Maximization"). If the decision maker is indifferent between her choices, I take no stand on how she resolves that indifference, and so the decision maker may flexibly randomize across choices whenever she is indifferent. Finally, the model-implied joint distribution of characteristics, choices and outcomes Q must match the observable joint distribution P after integrating out the private information ("Data Consistency"), linking the behavioral model to the decision maker's observed choices.

The key restriction of the expected utility maximization model is that only the observable characteristics $W \in \mathcal{W}$ directly affect the decision maker's utility function. Both the private information $V \in \mathcal{V}$ and the observable characteristics $X \in \mathcal{X}$ do not directly enter into the utility function, but are still payoff relevant since they enter the decision maker's information set and therefore affect her beliefs. In this sense, the decision maker's preferences satisfy an *exclusion restriction* on her private information $V \in \mathcal{V}$ and the observable characteristics X.

Since this is a substantive assumption on the decision maker's unknown preferences, it would seem difficult to justify such exclusion restrictions in practice.[6] I argue that the researcher may approach specifying such exclusion restrictions in three ways. First, in existing empirical research, such exclusion restrictions on the decision maker's preferences are already widespread. In pretrial release, it is common to assume that while judges may engage in taste-based discrimination based on defendant race, additional defendant characteristics such as their current charge or criminal record only affect beliefs about the likelihood of pretrial misconduct (e.g., see Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2020; Canay, Mogstad and Mountjoy, 2020; Hull, 2021). In medical testing, it is common to assume that a doctor's preferences are constant across patients, and observable patient characteristics only affect the doctor's beliefs about the probability of an underlying medical condition (e.g., see Abaluck et al., 2016; Mullainathan and Obermeyer, 2020; Chan, Gentzkow and Yu, 2020; Ribers and Ullrich, 2020). Therefore, the researcher may appeal to established, domain-specific modelling choices to guide the choice of characteristics $W \in \mathcal{W}$ vs. $X \in \mathcal{X}$. Second, the exclusion restriction may be motivated from a normative or legal perspective, and therefore summarizes social or legal restrictions on what observable characteristics ought not to directly enter preferences. In this view, Definition 3 should be interpreted as asking whether the decision maker's choices could have been generated by expected utility maximization behavior at *any* preferences that satisfy the normative exclusion restriction and private information. Finally, the researcher may conduct a sensitivity analysis on the choice of exclusion restriction, reporting how their analysis of expected utility maximization behavior varies as the choice of payoff-relevant characteristics $W \in \mathcal{W}$ varies. Such a sensitivity analysis summarizes how flexible the decision maker's preferences must be across observable characteristics in order to rationalize their observed choices.

Example: Pretrial Release The utility function specifies judge's relative payoffs from detaining a defendant that would not commit pretrial misconduct and releasing a defendant that would commit pretrial misconduct. These payoffs may vary based on some defendant characteristics W:

[6] Such an exclusion restriction on the decision maker's preferences echoes a long-standing debate in behavioral finance on whether observed variation in asset prices reflect violations of rational expectations or time variation in discount rates. See Campbell (2003), Cochrane (2011) for reviews and Augenblick and Lazarus (2020) for a recent contribution.

the judge may engage in tasted-based discrimination against black defendants (Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2020); be more lenient towards younger defendants (Stevenson and Doleac, 2019); or be more harsh towards defendants charged with violent crimes (Kleinberg et al., 2018). The judge's private information summarizes all additional information about defendants that is available to the judge at the time of her pretrial release decision, but not the researcher. For instance, the judge may glean useful information about defendants from court-room interactions, but these interactions are not recorded. ▲

If Definition 3 is satisfied, then the decision maker's implied beliefs about the outcome given the observable characteristics under the expected utility maximization model, denoted by $Q_{Y^*}(\cdot \mid w, x) \in \Delta(\mathcal{Y})$, are *accurate* in the following sense.

Lemma 2.1. *If the decision maker's choices are consistent with expected utility maximization behavior, then* $Q_{Y^*}(\cdot \mid w, x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w, x); \mathcal{B}_{c=0,w,x})$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$.

If the decision maker's choices are consistent with expected utility maximization behavior, then the decision maker is acting as-if she has beliefs about the outcome given the observable characteristics that lie in the identified set for the distribution of the outcome given the observable characteristics. Her implied beliefs must be "accurate" in this sense. This is an immediate consequence of the data consistency condition in Definition 3, and motivates the following definition of a systematic prediction mistake.

Definition 5. The decision maker is making *systematic prediction mistakes* based on the observable characteristics if her choices are inconsistent with expected utility maximization behavior, meaning $\mathcal{H}_P(U; \mathcal{B}_{c=0}) = \emptyset$.

The decision maker is making systematic prediction mistakes if there exists no configuration of preferences and private information such that her choices would maximize expected utility given some beliefs about the outcome conditional on the observable characteristics that lie in the identified set for the distribution of the outcome conditional on the observable characteristics. The decision maker's choices indicate that her implied beliefs about the outcome conditional on the observable characteristics are systematically incorrect in this sense.

Definitions 3-5 imply a weak notion of accurate beliefs, requiring only that her implied beliefs about the outcome given the observable characteristics, $Q_{Y^*}(\cdot \mid w, x)$ lie in the identified set for the distribution of the outcome given the observable characteristics, $\mathcal{H}_P(P_{Y^*}(\cdot \mid w, x); \mathcal{B}_{c=0,w,x})$. It does not require that the decision maker's implied beliefs be equal to the true marginal distribution $P(\cdot \mid w, x)$, and it also allows the decision maker to respond to "noisy" private information $V \in \mathcal{V}$ (e.g., see Kleinberg et al., 2018; Kahneman, 2019). In this sense, finding that the decision maker is making prediction mistakes according to Definition 5 is compelling evidence that the decision

maker is not optimizing given the available information – it means there are no beliefs about the outcome given the observable characteristics that lie in the identified set that could rationalize the decision maker's observed choices.

Importantly, this means that the interpretation of a prediction mistake in Definitions 3-5 is tied to the researcher-specified bounds on the missing data $\mathcal{B}_{c=0,w,x}$ in Assumption 2.1. Less informative bounds on the unobservable choice-dependent outcome probabilities imply that the expected utility maximization model places less restrictions on observed behavior as there are more candidate values of the missing choice-dependent outcome probabilities that may rationalize the decision maker's choices. Observed behavior that is consistent with expected utility maximization behavior at bounds $\mathcal{B}_{c=0,w,x}$ may, in fact, be inconsistent with expected utility maximization at alternative, tighter bounds $\widetilde{\mathcal{B}}_{c=0,w,x}$.[7] Definition 5 must therefore be interpreted as a prediction mistake that can be detected given the researcher-specified bounds on the missing data.

Remark 2.2 (Connection to Roy-Style Selection)**.** The expected utility maximization model relates to recent developments on Roy-style selection (Mourifie, Henry and Meango, 2019; Henry, Meango and Mourifie, 2020) and marginal outcome tests for taste-based discrimination (Bohren et al., 2020; Canay, Mogstad and Mountjoy, 2020; Gelbach, 2021; Hull, 2021). Defining the expected benefit functions $\Lambda_{c=0}(w,x,v) = \mathbb{E}_Q\left[U(0,Y^*;w) \mid W=w, X=x, V=v\right]$, $\Lambda_{c=1}(w,x,v) = \mathbb{E}_Q\left[U(1,Y^*;w) \mid W=w, X=x, V=v\right]$, the expected utility maximization model is a generalized Roy model that imposes that the observable characteristics $W \in \mathcal{W}$ enter into the utility function and affect beliefs, whereas the observable characteristics $X \in \mathcal{X}$ and private information $V \in \mathcal{V}$ only affect beliefs. The expected utility maximization model also takes no stand on how the decision maker resolves indifferences, and so it is an *incomplete* model of decision making.

2.3 Identification Result

The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ and values of the missing data that satisfy a series of revealed preference inequalities.

Theorem 2.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ and $\widetilde{P}_{Y^*}(\cdot \mid c=0,w,x) \in \mathcal{B}_{c=0,w,x}$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ satisfying*

$$\mathbb{E}_Q\left[U(c,Y^*;W) \mid C=c, W=w, X=x\right] \geq \mathbb{E}_Q\left[U(c',Y^*;W) \mid C=c, W=w, X=x\right]. \quad (1)$$

[7] Consider an extreme case in which $P(\cdot \mid w,x)$ is partially identified under bounds $\mathcal{B}_{c=0,w,x}$ but point identified under alternative bounds $\widetilde{\mathcal{B}}_{c=0,w,x}$. Under Definitions 3-5, a prediction mistake at bounds $\widetilde{\mathcal{B}}_{c=0,w,x}$ means that the decision maker's implied beliefs $Q_{Y^*}(\cdot \mid w,x)$ do not equal the point identified quantity $P(\cdot \mid w,x)$, yet a prediction mistake at bounds $\mathcal{B}_{c=0,w,x}$ means that the decision maker's implied beliefs $Q_{Y^*}(\cdot \mid w,x)$ do not lie in the identified set $\mathcal{H}_P(P(\cdot \mid w,x); \mathcal{B}_{c=0,w,x})$.

for all $c \in \{0,1\}$, $(w,x) \in \mathcal{W} \times \mathcal{X}$ *with* $P(C = c \mid W = w, X = x) > 0$ *and* $c' \neq c$, *where the joint distribution* $(W, X, C, Y^*) \sim Q$ *is given by*

$$Q(W = w, X = x, C = 1, Y^* = y^*) = P(W = w, X = x, C = 1, Y^* = y^*),$$
$$Q(W = w, X = x, C = 0, Y^* = y^*) = \widetilde{P}_{Y^*}(y^* \mid c = 0, w, x) P(W = w, X = x, C = 0).$$

Corollary 2.1. *The identified set of utility functions is*

$$\mathcal{H}_P(U; \mathcal{B}_{c=0}) = \left\{ U \in \mathcal{U} \colon \exists \widetilde{P}_{Y^*}(\cdot \mid c = 0, w, x) \in \mathcal{B}_{c=0,w,x} \text{ for all } (w,x) \in \mathcal{W} \times \mathcal{X} \text{ satisfying } (1) \right\}.$$

The revealed preference inequalities (1) require that there exists some value of the unobservable choice dependent outcome probabilities such that the decision maker's observed choices maximize expected utility at utility function $U \in \mathcal{U}$. If satisfied, then some private information $V \in \mathcal{V}$ can be constructed such that the decision maker's observed choices maximize expected utility given the observable characteristics and the constructed private information. If not, then there exists *no* private information such that the decision maker's choices maximize expected utility at those preferences. In this sense, Theorem 2.1 characterizes expected utility maximization behavior in a manner that is robust to weak assumptions on the decision maker's private information. These inequalities only involve the observable data and the bounds on the unobservable choice-dependent outcome probabilities, which will enable researchers to test whether these inequalities are satisfied as I show in Section 3.

Remark 2.3. Theorem 2.1 builds on the "no-improving action switches" inequalities in Caplin and Martin (2015), which characterize Bayesian expected utility maximization behavior in state-dependent stochastic choice data. There is no missing data problem in state-dependent stochastic choice data as the researcher observes the full joint distribution of choices and outcomes (typically because the researcher controls the entire decision-making environment in the laboratory). Theorem 2.1 characterizes expected utility maximization behavior in settings that suffer from a missing data problem, and is therefore applicable to empirical settings such as pretrial release, medical testing and hiring. While its proof is constructive, Theorem 2.1 can also be established indirectly through Bergemannn and Morris (2016)'s equivalence result between the set of Bayesian Nash Equilibrium and the set of Bayes Correlated Equilibrium in games of incomplete information, where the outcome Y^* is the state, the characteristics (W, X) are the signals and the private information V is the augmenting signal structure, and applying Data Consistency (Definition 3) on the equilibrium conditions.

2.4 Binary Screening Decisions

I explore the implications of Theorem 2.1 in a *binary screening decision*, which is an important empirical case where the outcome only takes two values $\mathcal{Y} = \{0, 1\}$ and covers the motivating applications of pretrial release and medical testing. The setting simplifies dramatically in a binary screening decision. First, the bounds on the unobservable choice-dependent outcome probabilities are now an interval for the conditional probability of $Y^* = 1$ given that the decision maker selected choice $C = 0$ with $\mathcal{B}_{c=0,w,x} = [\underline{P}(Y^* = 1 \mid C = 0, W = w, X = x), \overline{P}(Y^* = 1 \mid C = 0, W = w, X = x)]$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$. Second, it is without loss of generality to normalize two entries of the utility function, and so I normalize $U(0, 1; w) = 0, U(1, 0; w) = 0$ for all $w \in \mathcal{W}$.

I focus on testing whether the decision maker's choices are consistent with expected utility maximization behavior at strict preferences, meaning the decision maker strictly prefers a unique choice at each outcome under the expected utility maximization model. This is reasonable in empirical settings and rules out trivial cases such as complete indifference.

Assumption 2.2 (Strict Preferences)**.** The utility functions $U \in \mathcal{U}$ satisfy *strict preferences*, meaning $U(0, 0; w) < 0$ and $U(1, 1; w) < 0$ for all $w \in \mathcal{W}$.

Example: Pretrial Release The bounds $\mathcal{B}_{c=0,w,x} = [\underline{P}(Y^* = 1 \mid C = 0, W = w, X = x), \overline{P}(Y^* = 1 \mid C = 0, W = w, X = x)]$ specify lower and upper bounds respectively on the conditional pretrial misconduct rate among detained defendants. The utility function specifies the judge's cost of detaining a defendant that would not commit pretrial misconduct $U(0, 0; w)$ and releasing a defendant that would commit pretrial misconduct $U(1, 1; w)$. Focusing on strict preference utility functions imposes that no matter the observable characteristic of defendants, it is costly for the judge to either detain a defendant that would not commit pretrial misconduct or release a defendant that would commit pretrial misconduct.[8] ▲

By applying Theorem 2.1, I characterize the conditions under which the decision maker's choices in a binary screening decision are consistent with expected utility maximization behavior at some strict preference utility function and private information. For each $w \in \mathcal{W}$, define $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon P(C = 1 \mid W = w, X = x) > 0\}$ and $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon P(C = 0 \mid W = w, X = x) > 0\}$.

Theorem 2.2. *Consider a binary screening decision and assume* $P(Y^* = 1 \mid C = 1, W = w, X = x) < 1$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$ *with* $P(C = 1 \mid W = w, X = x) > 0$.

[8] In other empirical applications, it may be more economically reasonable to instead normalize $U(0, 0; w) = 0$, $U(1, 1; w) = 0$ and define strict preference utility functions to satisfy $U(0, 1; w) < 0, U(1, 0; w) < 0$ for all $w \in \mathcal{W}$. For example, in medical testing, it is more natural to assume that doctors would prefer to stress test a patient that did have a heart attack and would prefer to not stress test patients that did not have a heart attack.

The decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function if and only if for all $w \in \mathcal{W}$

$$\max_{x \in \mathcal{X}^1(w)} P(Y^* = 1 \mid C = 1, W = w, X = x) \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}(Y^* = 1 \mid C = 0, W = w, X = x).$$

Otherwise, $\mathcal{H}_P(U; \mathcal{B}_{c=0}) = \emptyset$*, and the decision maker is making prediction mistakes based on the observable characteristics.*

Corollary 2.2. *The identified set of strict preference utility functions* $\mathcal{H}_P(U; \mathcal{B}_{c=0})$ *is equal to the set of all utility functions satisfying for all* $w \in \mathcal{W}$*,* $U(0,0;w) < 0$*,* $U(1,1;w) < 0$ *and*

$$\max_{x \in \mathcal{X}^1(w)} P(Y^* = 1 \mid C = 1, W = w, X = x) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)},$$
$$\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}(Y^* = 1 \mid C = 0, W = w, X = x).$$

The decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function if and only if the decision maker is acting *as-if* she applies a threshold rule that depends on the characteristics W. The threshold equals $\frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)}$ and summarizes the relative costs of "ex-post errors" – i.e., the relative cost of choosing $C = 0$ for an individual with outcome $Y^* = 0$ and choosing $C = 1$ for an individual with outcome $Y^* = 1$. If the conditions in Theorem 2.2 are violated, there exists no strict preference utility function nor private information such that the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs.

Theorem 2.2 highlights the necessity of placing an exclusion restriction on which observable characteristics directly affect the decision maker's utility function. If all observable characteristics directly affect the decision maker's utility function (i.e., $\mathcal{X} = \emptyset$), then the decision maker's choices are consistent with expected utility maximization behavior whenever the researcher assumes the decision maker observes private information that is predictive of the outcome.

Corollary 2.3. *Under the same conditions as Theorem 2.2, suppose* $\mathcal{X} = \emptyset$ *and all observable characteristics directly affect the decision maker's utility function.*

If $P(Y^* = 1 \mid C = 1, W = w) \leq \overline{P}(Y^* = 1 \mid C = 0, W = w)$ *for all* $w \in \mathcal{W}$*, then the decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function.*

This negative result arises because a characteristic-dependent threshold can always be constructed that rationalizes the decision maker's observed choices if the probability of $Y^* = 1$ given $C = 0$ is at least as large as the observed probability of $Y^* = 1$ given $C = 1$ for all characteristics.

Therefore, imposing an exclusion restriction that some observable characteristics do not directly enter into the utility function is necessary for testing whether the decision maker makes prediction mistakes if we suspect the decision maker observes useful private information.

Unfortunately, imposing such an exclusion restriction is not alone sufficient to restore the testability of expected utility maximization behavior, and therefore prediction mistakes. The researcher must still address the missing data problem.

Corollary 2.4. *Under the same conditions as Theorem 2.2, if* $\overline{P}(Y^* = 1 \mid C = 0, W = w, X = x) = 1$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$*, then the decision maker's choices are consistent with expected utility maximization behavior at some strict preferences.*

Without assumptions or a research design to generate informative bounds on the unobservable choice-dependent outcome probabilities, the decision maker's choices are always consistent with maximizing expected utility at some strict preference utility function. If the unobservable choice-dependent outcome probabilities may take any value, then the decision maker's choices may be rationalized by the extreme case in which the decision maker's private information is perfectly predictive of the unknown outcome (i.e., $\overline{P}(Y^* = 1 \mid C = 0, W = w, X = x) = 1$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$).

Corollaries 2.3-2.4 highlight that testing expected utility maximization behavior, and therefore prediction mistakes, requires *both* behavioral assumptions on which observable characteristics may directly affect the decision maker's preferences *and* econometric assumptions that generate informative bounds on the unobservable choice-dependent outcome probabilities. Under such assumptions, Theorem 2.2 provides interpretable conditions to identify prediction mistakes. At any fixed $w \in \mathcal{W}$, does there exist some $x \in \mathcal{X}$ such that the largest possible probability of $Y^* = 1$ given $C = 0$ is strictly lower than the observed probability of outcome $Y^* = 1$ given $C = 1$ at some other $x' \in \mathcal{X}$? If so, then the decision maker cannot be maximizing expected utility as the decision maker could do strictly better by flipping her choices across x, x'. She could do strictly better by raising her probability of selecting choice $C = 0$ at x' and lowering her probability of selecting choice $C = 1$ at x. These "misranking" arguments are necessary and sufficient to test the joint null hypothesis that the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs and her preferences satisfy the conjectured exclusion restriction.

Example: Pretrial Release The conditions in Theorem 2.2 require that, holding fixed the defendant characteristics that directly affect the judge's preferences, all detained defendant must have a higher worst-case probability of committing pretrial misconduct than any released defendant. More concretely, suppose the researcher assumes that the judge may engage in taste-based discrimination based on the defendant race W, but the judge's preferences are unaffected by remaining defendant characteristics X. Imposing such an exclusion restriction generates testable

restrictions on the judge's release decisions. Within defendants of the same race, does there exist some group of released defendants with a higher observed pretrial misconduct rate than the worst-case pretrial misconduct rate of some group of detained defendants? If so, the judge is misranking the probability of pretrial misconduct between these two groups of defendants, and could do strictly better off by flipping her choices across these groups. The judge's choices are inconsistent with expected utility maximization behavior at preferences that only depend on defendant race and accurate beliefs given defendant characteristics as well as some private information. ▲

Remark 2.4 (Connection to Marginal Outcome Tests and Infra-Marginality)**.** In a binary screening decision, my analysis complements recent results in Canay, Mogstad and Mountjoy (2020), Gelbach (2021), and Hull (2021), which use an extended Roy model to explore the validity of marginal outcome tests for taste-based discrimination. Canay, Mogstad and Mountjoy (2020), Gelbach (2021), and Hull (2021) exploit the continuity of private information to derive testable implications of the model in terms of underlying marginal treatment effect functions, which requires that the researcher identify the conditional expectation of the outcome at each possible "marginal" decision. In contrast, the characterization in Theorem 2.2 involves only "inframarginal" outcomes. As I discuss in the next section, common sources of quasi-experimental variation naturally lend themselves to bounds on these inframarginal outcomes without requiring additional behavioral assumptions such as monotonicity or additional functional form restrictions needed to estimate marginal treatment effect curves as discussed in Hull (2021).

3 Testing Expected Utility Maximization Behavior

In this section, I show how the researcher may test whether the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs based on the identification results in Section 2. First, I discuss how the researcher may construct informative bounds on the unobservable choice-dependent outcome probabilities by leveraging a randomly assigned instrument. Second, given some constructed bounds, I show how testing the revealed preference inequalities is equivalent to testing many moment inequalities. This section focuses on binary screening decisions, and I provide general results for the many outcomes setting in Appendix B.

3.1 Constructing Bounds on the Missing Data with an Instrument

Suppose there exists a randomly assigned instrument that generates variation in the decision maker's choice probabilities. How may such an instrument be used to construct informative bounds on the unobservable choice-dependent outcome probabilities? Instruments commonly arise, for example, through the random assignment of decision makers. Judges are randomly assigned to defendants in pretrial release (e.g., Dobbie, Goldin and Yang, 2018; Arnold, Dobbie and Yang, 2018; Kleinberg

et al., 2018; Arnold, Dobbie and Hull, 2020) and doctors may be randomly assigned to patients in medical testing (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2020). While the random assignment of decision makers is a common example, it is not the only possibility. For instance, Mullainathan and Obermeyer (2020) argue that doctors are less likely to conduct stress tests for a heart attack on Fridays and Saturdays due to weekend staffing constraints, even though patients that arrive on these days are no less risky. The random introduction of or changes to recommended guidelines may also affect decision makers' choices (e.g., see Albright, 2019; Abaluck et al., 2020).

I now assume that the researcher additionally observes a randomly assigned instrument $Z \in \mathcal{Z}$.

Assumption 3.1 (Random Instrument)**.** Let $Z \in \mathcal{Z}$ be a finite support instrument, and the joint distribution $(W, X, Z, C, Y^*) \sim P$ satisfies $(W, X, Y^*) \perp\!\!\!\perp Z$ under P and $P(W = w, X = x, Z = z) > 0$ for all $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$.

The researcher observes the joint distribution $(W, X, Z, C, Y) \sim P$, where $Y = C \cdot Y^*$ as before. Under Assumption 3.1, the unobservable choice-dependent outcome probabilities are partially identified, denoting their sharp identified sets as $\mathcal{H}_P(P_{Y^*}(\cdot \mid c = 0, w, x, z))$.

Proposition 3.1. *Suppose Assumption 3.1 holds and consider a binary screening decision. Then, for any* $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$ *with* $P(C = 0 \mid W = w, X = x, Z = z) > 0$, $\mathcal{H}_P(P_{Y^*}(\cdot \mid c = 0, w, x, z)) = \left[\underline{P}(Y^* = 1 \mid C = 0, W = w, X = x, Z = z), \overline{P}(Y^* = 1 \mid C = 0, W = w, X = x, Z = z)\right]$, *where*

$$\underline{P}(Y^* = 1 \mid C = 0, W = w, X = x, Z = z) =$$

$$\max\left\{\frac{\underline{P}(Y^* = 1 \mid W = w, X = x) - P(C = 1, Y^* = 1 \mid W = w, X = x, Z = z)}{P(C = 0 \mid W = w, X = x, Z = z)}, 0\right\},$$

$$\overline{P}(Y^* = 1 \mid C = 0, W = w, X = x, Z = z) =$$

$$\min\left\{\frac{\overline{P}(Y^* = 1 \mid W = w, X = x) - P(C = 1, Y^* = 1 \mid W = w, X = x, Z = z)}{P(C = 0 \mid W = w, X = x, Z = z)}, 1\right\},$$

and $\underline{P}(Y^* = 1 \mid W = w, X = x) = \max_{\tilde{z} \in \mathcal{Z}}\{P(C = 1, Y^* = 1 \mid W = w, X = x, Z = \tilde{z})\}$, $\overline{P}(Y^* = 1 \mid W = w, X = x) = \min_{\tilde{z} \in \mathcal{Z}}\{P(C = 0 \mid W = w, X = x, Z = \tilde{z}) + P(C = 1, Y^* = 1 \mid W = w, X = x, Z = \tilde{z})\}$.

Therefore, within a fixed value $z \in \mathcal{Z}$, the researcher may apply the identification results derived in Section 2 by defining $\mathcal{B}_{c=0,w,x} = \mathcal{H}_P(P_{Y^*}(\cdot \mid c = 0, w, x, z))$ under Assumption 3.1. Appendix B.1 extends these bounds to allow for the instrument to be quasi-randomly assigned conditional on some additional characteristics.

An active literature examines the behavioral interpretation of instrumental variable research designs in cases where decision makers are randomly assigned to decisions, focusing on the concern that judges designs may violate the classic monotonicity assumption in Imbens and Angrist

(1994); Angrist, Imbens and Rubin (1996). In pretrial release, monotonicity requires that if a strict judge detains more defendants than a lenient judge, then every defendant detained by the strict judge would also be detained by the lenient judge. de Chaisemartin (2017); Frandsen, Lefgren and Leslie (2019); Chan, Gentzkow and Yu (2020) develop weaker notions of monotonicity for these settings that still preserve a causal interpretation to linear instrumental variables estimators. I emphasize that I impose no form of monotonicity (or its relaxations) in deriving the bounds in Proposition 3.1.

Remark 3.1 (Contraction across decision makers)**.** Lakkaraju et al. (2017) and Kleinberg et al. (2018) develop "contraction," which is a procedure that uses the random assignment of decision makers to evaluate a statistical decision rule $\tilde{C}$ by imputing its true positive rate $P(Y^* = 1 \mid \tilde{C} = 1)$. In contrast, Proposition 3.1 uses an instrument to construct bounds on a decision maker's unobservable choice-dependent outcome probabilities, $P(Y^* = 1 \mid C = 0, W = w, X = x)$. As shown in Section 2, bounds on a decision maker's unobservable choice-dependent outcome probabilities are necessary in order to test for detectable prediction mistakes.

Remark 3.2 (Other empirical strategies for constructing bounds)**.** In empirical applications, there may be no instrument available, and so I discuss two additional empirical strategies in Supplement E. First, researchers may bound the unobservable choice-dependent outcome probabilities using the observable choice-dependent outcome probabilities through what I call "direct imputation." This is analogous to sensitivity analysis techniques in causal inference such as Rosenbaum (2002). Second, the researcher may use a "proxy outcome," which does not suffer the missing data problem and is correlated with the outcome.

3.1.1 Behavioral Interpretation

Assumption 3.1 imposes an important behavioral restriction on the expected utility maximization model, requiring that the decision maker's implied beliefs about the unknown outcome do not depend on the instrument.

Proposition 3.2. *Suppose Assumption 3.1 holds. If the decision maker's choices are consistent with expected utility maximization behavior at some utility function U and joint distribution $(W, X, Z, V, C, Y^*) \sim Q$, then $Y^* \perp\!\!\!\perp Z \mid W, X$ under Q.*

Requiring that the decision maker's beliefs about the unknown outcome be accurate imposes that the instrument cannot affect their beliefs. Aside from this restriction on implied beliefs, Assumption 3.1 places no further behavioral restrictions on the model. The instrument may affect preferences and private information. Consider pretrial release decisions in which the instrument arises

through the random assignment of judges, and so $Z \in \mathcal{Z}$ refers to a judge identifier. Proposition 3.2 then states that if all judges make choices as-if they are maximizing expected utility at accurate beliefs given defendant characteristics, then all judges must have the same beliefs about the probability of pretrial misconduct given defendant characteristics since the random assignment of judges to defendants implies that a judge's identity is independent of whether a defendant will commit pretrial misconduct. However, judges may still differ from one another in their preferences and private information. In this sense, the expected utility maximization model still allows for rich heterogeneity in behavior across judges.[9]

3.2 Testing Expected Utility Maximization Behavior Reduces to Testing Many Moment Inequalities

Testing whether the decision maker's choices in a binary screening decision are consistent with expected utility maximization behavior at strict preferences (Theorem 2.2) reduces to testing many moment inequalities.

Proposition 3.3. *Consider a binary screening decisions. Suppose Assumption 3.1 holds, and* $0 < P(C = 1 \mid W = w, X = x, Z = z) < 1$ *for all* $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$.

The decision maker's choices at $z \in \mathcal{Z}$ *are consistent with expected utility maximization behavior at some strict preference utility function if and only if for all* $w \in \mathcal{W}$, *pairs* $x, \tilde{x} \in \mathcal{X}$ *and* $\tilde{z} \in \mathcal{Z}$

$$P(Y^* = 1 \mid C = 1, W = w, X = x, Z = z) - \overline{P}_{\tilde{z}}(Y^* = 1 \mid C = 0, W = w, X = \tilde{x}, Z = z) \leq 0,$$

where $\overline{P}_{\tilde{z}}(Y^* = 1 \mid C = 0, W = w, X = x, Z = z) = \frac{P(C=0|W=w,X=x,Z=\tilde{z})+P(C=1,Y^*=1|W=w,X=x,Z=\tilde{z})}{P(C=0|W=w,X=x,Z=z)} - \frac{P(C=1,Y^*=1|W=w,X=x,Z=z)}{P(C=0|W=w,X=x,Z=z)}$.

The number of moment inequalities is equal to $d_w \cdot d_x^2 \cdot (d_z - 1)$, and grows with the number of support points of the characteristics and instruments. In empirical applications, this will be quite large since the characteristics of individuals are extremely rich. Consequently, testing the revealed preference inequalities in a binary screening decision directly over the underlying characteristics may require using moment inequality procedures that are valid in high-dimensional settings such as Chernozhukov, Chetverikov and Kato (2019) and Bai, Santos and Shaikh (2021).

The number of moment inequalities may be reduced by testing a set of implied revealed preference inequalities over any partition of the excluded characteristics. For each $w \in \mathcal{W}$, define $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ to be some function that partitions the support of the excluded characteristics $x \in \mathcal{X}$ into level sets $\{x \colon D_w(x) = d\}$. By iterated expectations, if the decision maker's

[9] Allowing decision makers to vary in their private information has been recently described as allowing decision makers to have "varying predictive skill" in Chan, Gentzkow and Yu (2020); Arnold, Dobbie and Hull (2020).

choices are consistent with expected utility maximization behavior at some utility function $U \in \mathcal{U}$, then her choices must satisfy *implied* revealed preference inequalities.

Corollary 3.1. *Consider a binary screening decision, and assume* $P(Y^* = 1 \mid C = 1, W = w, X = x) < 1$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$ *with* $P(C = 1 \mid W = w, X = x) > 0$. *If the decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function, then for all* $w \in \mathcal{W}$

$$\max_{d \in \mathcal{D}^1(w)} P(Y^* = 1 \mid C = 1, W = w, D_w(X) = d) \leq \min_{x \in \mathcal{D}^0(w)} \overline{P}(Y^* = 1 \mid C = 0, W = w, D_w(X) = d),$$

where $\mathcal{D}^1(w) := \{d \colon P(C = 1 \mid W = w, D_w(X) = d) > 0\}$ *and* $\mathcal{D}^0(w) := \{d \colon P(C = 0 \mid W = w, D_w(X) = d) > 0\}$.

In practice, this can drastically reduce the number of moment inequalities that must be tested. If $N_d \ll d_x$, researchers may test the implied moment inequalities in Corollary 3.1 using procedures that are valid in low-dimensional settings, which is a mature literature in econometrics – see, for example, the reviews in Canay and Shaikh (2017), Ho and Rosen (2017) and Molinari (2020).

A natural choice is to construct the partitioning functions $D_w(\cdot)$ using supervised machine learning methods to predict the outcome on a set of held-out decisions. The researcher may construct an estimated prediction function $\hat{f} \colon \mathcal{W} \times \mathcal{X} \to [0, 1]$ to predict the outcome. Given the estimated prediction function, the researcher may define $D_w(x)$ by binning the characteristics X into percentiles of predicted risk within each value $w \in \mathcal{W}$. The implied revealed preference inequalities then search for misrankings in the decision maker's choices across percentiles of predicted risk. Such an empirical strategy is close to existing empirical practice as it is already common to compare the choices of decision makers against the recommended choices of an estimated decision rule (e.g. Kleinberg et al., 2018; Chouldechova et al., 2018; Mullainathan and Obermeyer, 2020; Ribers and Ullrich, 2019, 2020). These results provide, to the best of my knowledge, the first formal procedure for using such estimated prediction functions to formally test whether the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs.

4 Identifying Prediction Mistakes based on Characteristics

So far, I have shown that researchers can test whether a decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs about the outcome, and therefore whether the decision maker is making prediction mistakes. By modifying the expected utility maximization model and the revealed preference inequalities, I next show that researchers can further investigate the types of prediction mistakes that are being made.

4.1 What Types of Prediction Mistakes are Being Made?

If the decision maker's choices are inconsistent with expected utility at accurate beliefs about the outcome under some specified exclusion restriction on their preferences, what types of prediction mistakes are being made? By modifying the expected utility maximization model, the researcher may construct bounds on an interpretable parameter that summarizes the extent to which the decision maker's beliefs about the outcome underreact or overreact to variation in the outcome based on the characteristics, and therefore provide a summary measure of the types of prediction mistakes made by the decision maker.

4.1.1 Bounding Prediction Mistakes in Binary Screening Decisions

Recall that the definition of expected utility maximization behavior (Definition 3) implied that the decision maker acted as-if her implied beliefs about the outcome given the observable characteristics were accurate (Lemma 2.1). As a result, the revealed preference inequalities may be violated if the decision maker acted as-if she maximized expected utility based on *inaccurate* beliefs about the outcome given the characteristics, meaning that the decision maker's implied beliefs do not lie in the identified set of the marginal distribution of the outcome given the characteristics $\mathcal{H}_P(P_{Y^*}(\cdot \mid w, x))$. This is a common behavioral hypothesis in empirical applications. For example, researchers conjecture that judges may systematically mis-predict failure to appear risk based on defendant characteristics, and the same concern arises in analyses of medical testing and treatment decisions.[10]

To investigate the possibility that the decision maker's choices may maximize expected utility at inaccurate beliefs given the observable characteristics, I modify the "Data Consistency" condition in the definition of expected utility maximization behavior.

Definition 6. The decision maker's choices are *consistent with expected utility maximization behavior at inaccurate beliefs* if there exists some utility function $U \in \mathcal{U}$ and joint distribution $(W, X, V, C, Y^*) \sim Q$ satisfying (i) Information Set, (ii) Expected Utility Maximization, and

iii. Data Consistency with Inaccurate Beliefs: For all $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists $\widetilde{P}_{Y^*}(\cdot \mid c = 0, w, x) \in \mathcal{B}_{c=0,w,x}$ such that for all $y^* \in \mathcal{Y}$

$$P(C = 1, Y^* = y^*, W = w, X = x) =$$

$$Q(C = 1 \mid Y^* = y^*, W = w, X = x)\widetilde{P}_{Y^*}(y^* \mid w, x)Q(W = w, X = x),$$

[10] Kleinberg et al. (2018) write, "a primary source of error is that all quintiles of judges misuse the signal available in defendant characteristics available in our data" (pg. 282-283). In the medical treatment setting, Currie and Macleod (2017) write, "We are concerned with doctors, who for a variety of possible reasons, do not make the best use of the publicly available information at their disposal to make good decisions" (pg. 5).

$$\widetilde{P}_{Y^*}(y^* \mid c=0,w,x)P(C=0,W=w,X=x) =$$

$$Q(C=0 \mid Y^*=y^*,W=w,X=x)\widetilde{P}_{Y^*}(y^* \mid w,x)Q(W=w,X=x),$$

where $\widetilde{P}_{Y^*}(y^* \mid w,x) = P(C=1,Y^*=y^* \mid W=w,X=x) + \widetilde{P}_{Y^*}(y^* \mid c=0,w,x)P(C=0 \mid W=w,X=x)$.

Definition 6 requires that the joint distribution $(W,X,V,C,Y^*) \sim Q$ under the expected utility maximization model only matches the observable joint distribution P if we replace the decision maker's model-implied beliefs given the characteristics, $Q_{Y^*}(\cdot \mid w,x)$, with some marginal distribution of the outcome given the characteristics that lies in the identified set, $\widetilde{P}_{Y^*}(\cdot \mid w,x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w,x)[\mathcal{B}_{c=0,w,x}})$. Intuitively, this imposes that the decision maker correctly specifies the likelihood of their private information $V \mid Y^*,W,X$, and so prediction mistake arises only arise from her misspecified "prior" beliefs about the outcome given the characteristics (i.e., the conditional distribution $Y^* \mid W,X$). Therefore, her prediction mistakes are driven solely by incorrectly specifying the marginal distribution of the outcome given the characteristics.

As a stepping stone to investigating the decision maker's prediction mistakes, I derive an identified set for a reweighed version of decision maker's strict preference utility function in a binary screening decision.

Theorem 4.1. *Consider a binary screening decision. Assume* $0 < P(Y^*=1 \mid W=w,X=x) < 1$ *for all* $P_{Y^*}(\cdot \mid w,x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w,x);\mathcal{B}_{c=0,w,x})$ *and* $0 < P(C=1 \mid W=w,X=x) < 1$ *for all* $(w,x) \in \mathcal{W} \times \mathcal{X}$.

Suppose the decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs and some strict preference utility function. Then, there exists non-negative weights $\omega(y^*;w,x) \geq 0$ *satisfying for all* $(w,x) \in \mathcal{W} \times \mathcal{X}$

$$P(Y^*=1 \mid C=1,W=w,X=x) \leq \frac{\omega(0;w,x)U(0,0;w)}{\omega(0;w,x)U(0,0;w)+\omega(1;w,x)U(1,1;w)} \tag{2}$$
$$\frac{\omega(0;w,x)U(0,0;w)}{\omega(0;w,x)U(0,0;w)+\omega(1;w,x)U(1,1;w)} \leq \overline{P}(Y^*=1 \mid C=0,W=w,X=x),$$

where $\omega(y^*;w,x) = Q(Y^*=y^* \mid W=w,X=x)/\widetilde{P}(Y^*=y^* \mid W=w,X=x)$ *and* $Q(Y^*=y^* \mid W=w,X=x)$, $\widetilde{P}(Y^*=y^* \mid W=w,X=x)$ *are defined in Definition 6.*

To prove this result, I show that a modified set of revealed preference inequalities characterize whether the decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs. These modified inequalities ask whether the decision maker's observed choices satisfy revealed preference inequalities at some reweighed utility function, where the weights are the likelihood ratio of the decision maker's implied beliefs given the characteristics relative to the

true conditional distribution of the outcome given the characteristics. Theorem 4.1 follows by re-arranging the modified revealed preference inequalities.

Theorem 4.1 shows that allowing the decision maker to have inaccurate beliefs about the outcome given the characteristics is equivalent to allowing the decision maker's reweighed utility function to depend on the characteristics $x \in \mathcal{X}$ only through the weights $\omega(y^*; w, x)$, which summarize the extent to which the decision maker mispredicts the outcome. Behavior that is consistent with expected utility maximization behavior at inaccurate beliefs is then equivalent to behavior that follows a threshold rule that depends on all characteristics in this manner.

This intuition may be exploited to derive an identified set on the extent to which the decision maker overreacts or underreacts to variation in the characteristics. Define $\delta(w,x) := \frac{Q(Y^*=1|W=w,X=x)/Q(Y^*=0|W=w,X=x)}{\widetilde{P}(Y^*=1|W=w,X=x)/\widetilde{P}(Y^*=0|W=w,X=x)}$ to be the relative odds ratio of the unknown outcome under the decision maker's implied beliefs relative to the true conditional distribution, and $\tau(w,x) := \frac{\omega(0;w,x)U(0,0;w)}{\omega(0;w,x)U(0,0;w)+\omega(1;w,x)U(1,1;w)}$ to be the decision maker's reweighed utility threshold. If the reweighed utility threshold were known, then the decision maker's implied prediction mistake can be backed out.

Corollary 4.1. *Under the same conditions as Theorem 4.1,*

$$\frac{(1-\tau(w,x))/\tau(w,x)}{(1-\tau(w,x'))/\tau(w,x')} = \frac{\delta(w,x)}{\delta(w,x')} \tag{3}$$

or any $w \in \mathcal{W}$, $x, x' \in \mathcal{X}$.

The ratio $\frac{\delta(w,x)}{\delta(w,x')}$ summarizes the extent to which the decision maker's implied beliefs about the outcome overreact or underreact to variation in the characteristics relative to the true conditional distribution. Notice that it may be rewritten as the ratio of $\frac{Q(Y^*=1|W=w,X=x)/Q(Y^*=0|W=w,X=x)}{Q(Y^*=1|W=w,X=x')/Q(Y^*=0|W=w,X=x')}$ to $\frac{\widetilde{P}(Y^*=1|W=w,X=x)/\widetilde{P}(Y^*=0|W=w,X=x)}{\widetilde{P}(Y^*=1|W=w,X=x')/\widetilde{P}(Y^*=0|W=w,X=x')}$, where the first term summarizes how the odds ratio of $Y^* = 1$ relative to $Y^* = 0$ varies across (w,x) and (w,x') under the decision maker's implied beliefs and the second term summarizes how the true odds ratio of the outcome varies across the same values. If $\frac{\delta(w,x)}{\delta(w,x')}$ is strictly less than one, then the decision maker's implied beliefs react less to variation across (w,x) and (w,x') than the true distribution, and her implied beliefs are *underreacting* across these characteristics. Analogously if $\frac{\delta(w,x)}{\delta(w,x')}$ is strictly greater than one, then the decision maker's implied beliefs are *overreacting*.

Since Theorem 4.1 provides an identified set for the reweighted utility thresholds, an identified set for the implied prediction mistake $\frac{\delta(w,x)}{\delta(w,x')}$ can in turn be constructed by computing the ratio (3) for each pair $\tau(w,x), \tau(w,x')$ that satisfies (2). Therefore, provided the decision maker's choice are consistent with expected utility maximization behavior at inaccurate beliefs and some strict preference utility function, bounds may be constructed on the decision maker's implied prediction

mistakes. These bounds are not sharp as there may exist thresholds $\tau(w, x)$ that satisfy (2) but are not consistent with expected utility maximization behavior at inaccurate beliefs.

This result stands in contrast to Proposition 1 in Martin and Marx (2021), which shows that utilities and prior beliefs are not separately identified in state-dependent stochastic choice data (see also Arnold, Dobbie and Yang (2018) and Bohren et al. (2020)). Their non-identification result arises because the authors focus on settings in which there are no additional observable characteristics of decisions beyond those which are directly affect utility. In contrast, in empirical settings such as pretrial release and medical testing, researchers commonly assume that there exists defendant characteristics that only enter into judges' information sets. Variation in the decision maker's choices and outcomes across observable characteristics that do not directly affect preferences are informative about the decision maker's beliefs, and therefore such preference exclusion restrictions are sufficient to partially identify a the types of systematic prediction mistakes that are being made by the decision maker.

Remark 4.1. After applying the dimension reduction strategy in Section 3.2, the implied prediction mistake $\delta(w, d)/\delta(w, d')$ across values $D_w(X) = d, D_w(X) = d'$ now measures how the decision maker's implied beliefs of her own ex-post mistakes varies relative to the true probability of ex-post mistakes across values $D_w(X) = d, D_w(X) = d'$. This is still an informative summary of the decision maker's prediction mistakes. See Appendix A.1 for details.

5 Conclusion

This paper develops an econometric framework for testing whether a decision maker makes prediction mistakes in high stakes settings such as pretrial release and medical testing. I characterized the implications of expected utility maximization behavior, where the decision maker maximizes some utility function at accurate beliefs about the outcome given the observable characteristics of each decision as well as some private information. These implications are testable, and I show how researchers may leverage existing econometric procedures and supervised machine learning methods to develop computationally tractable tests.

More broadly, this paper highlights that prediction policy problems such as pretrial release and medical testing can serve as rich laboratories for behavioral and policy analysis (Kleinberg et al., 2015). I provided a first step by exploring the empirical implications of a canonical model of decision making under uncertainty, expected utility maximization, in these settings. An exciting avenue for future research is to explore the testable implications of alternative behavioral models such as rational inattention (e.g., Sims, 2003; Gabaix, 2014; Caplin and Dean, 2015) as well as salience and stereotypes (e.g., Gennaioli and Shleifer, 2010; Bordalo et al., 2016). What are the testable implications of these behavioral models in leading prediction problems? Furthermore, the design of algorithmic decision system requires both identifying systematic prediction mistakes when they

occur and empirically characterizing the tradeoff between delegating decisions to a decision maker that observes useful private information but may be misaligned with the policymaker's objectives. Exploiting the full potential of these empirical settings is an exciting agenda at the intersection of economic theory, machine learning, and microeconometrics.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. 2016. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care." *American Economic Review*, 106(12): 3730–3764.

Abaluck, Jason, Leila Agha, David C. Chan, Daniel Singer, and Diana Zhu. 2020. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." NBER Working Paper No. 27467.

Ahn, David S., and Todd Sarver. 2013. "Preference for Flexibility and Random Choice." *Econometrica*, 81(1): 341–361.

Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions."

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes. 2019. "Inference for Linear Conditional Moment Inequalities." NBER Working Paper No. 26374.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91: 444–455.

Arnold, David, Will Dobbie, and Crystal Yang. 2018. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Arnold, David, Will Dobbie, and Peter Hull. 2020. "Measuring Racial Discrimination in Bail Decisions." NBER Working Paper No. 26999.

Augenblick, Ned, and Eben Lazarus. 2020. "Restrictions on Asset-Price Movements Under Rational Expectations: Theory and Evidence."

Autor, David H., and David Scarborough. 2008. "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments." *The Quarterly Journal of Economics*, 123(1): 219–277.

Bai, Yuehao, Andres Santos, and Azeem M. Shaikh. 2021. "A Practical Method for Testing Many Moment Inequalities." *Journal of Business Economics and Statistics*.

Beaulieu-Jones, Brett, Samuel G. Finlayson, Corey Chivers, Irene Chen, Matthew McDermott, Jaz Kandola, Adrian V. Dalca, Andrew Beam, Madalina Fiterau, and Tristan Naumann. 2019. "Trends and Focus of Machine Learning Applications for Health Research." *JAMA Network Open*, 2(10): e1914051–e1914051.

Belloni, Alexandre, Federico Bugni, and Victor Chernozhukov. 2018. "Subvector Inference in Partially Identified Models with Many Moment Inequalities." arXiv preprint, arXiv:1806.11466.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris. 2019. "Counterfactuals with Latent Information."

Bergemannn, Dirk, and Stephen Morris. 2013. "Robust Predictions in Games with Incomplete Information." *Econometrica*, 81(4): 1251–1308.

Bergemannn, Dirk, and Stephen Morris. 2016. "Bayes correlated equilibrium and the comparison of information structures in games." *Theoretical Economics*, 11: 487–522.

Bergemannn, Dirk, and Stephen Morris. 2019. "Information Design: A Unified Perspective." *Journal of Economic Literature*, 57(1): 44–95.

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies*, 13(1): 94–115.

Blattner, Laura, and Scott T. Nelson. 2021. "How Costly is Noise?" arXiv preprint, arXiv:arXiv:2105.07554.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2020. "Inaccurate Statistical Discrimination: An Identification Problem." NBER Working Paper Series No. 25935.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics*, 131(4): 1753–1794.

Bugni, Federico A., Ivan A. Canay, and XiaoXia Shi. 2015. "Specification tests for partially identified models defined by moment inequalities." *Journal of Econometrics*, 185: 259–282.

Camerer, Colin F. 2019. "Artificial Intelligence and Behavioral Economics." In *The Economics of Artificial Intelligence: An Agenda*. , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb, 587–608. University of Chicago Press.

Camerer, Colin. F., and Eric J. Johnson. 1997. "The Process-Performance Paradox in Expert Judgement." In *Research on Judgment and Decision Making: Currents, Connections, and Controversies*. , ed. W. M. Goldstein and R. M. Hogarth. New York: Cambridge University Press.

Campbell, John Y. 2003. "Chapter 13 Consumption-based asset pricing." In *Financial Markets and Asset Pricing*. Vol. 1 of *Handbook of the Economics of Finance*, 803–887. Elsevier.

Campbell, John Y., and Robert J. Shiller. 1987. "Cointegration and Tests of Present Value Models." *The Journal of Political Economy*, 95(5): 1062–1088.

Canay, Ivan A., and Azeem M. Shaikh. 2017. "Practical and Theoretical Advances in Inference for Partially Identified Models." *Advances in Economics and Econometrics: Eleventh World Congress*, , ed. Bo Honoré, Ariel Pakes, Monika Piazzesi and Larry Samuelson Vol. 2, 271–306. Cambridge University Press.

Canay, Ivan, Magne Mogstad, and Jack Mountjoy. 2020. "On the Use of Outcome Tests for Detecting Bias in Decision Making." NBER Working Paper No. 27802.

Caplin, Andrew. 2016. "Measuring and Modeling Attention." *Annual Review of Economics*, 8: 379–403.

Caplin, Andrew, and Daniel Martin. 2015. "A Testable Theory of Imperfect Perception." *Economic Journal*, 125: 184–202.

Caplin, Andrew, and Daniel Martin. 2020. "Framing, Information and Welfare." NBER Working Paper No. 27265.

Caplin, Andrew, and Mark Dean. 2015. "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review*, 105(7): 2183–2203.

Caplin, Andrew, Dànіel Csaba, John Leahy, and Oded Nov. 2020. "Rational Inattention, Competitive Supply, and Psychometrics." *The Quarterly Journal of Economics*, 135(3): 1681–1724.

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. "Search and Satisficing." *American Economic Review*, 101(7): 2899–2922.

Carroll, Christopher D. 2003. "Macroeconomic Expectations of Households and Professional Forecasters." *The Quarterly Journal of Economics*, 118(1): 269–298.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review*, 106(5): 124–127.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2020. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." NBER Working Paper No. 26467.

Chandra, Amitabh, and Douglas O. Staiger. 2007. "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks." *Journal of Political Economy*, 115(1): 103–140.

Chandra, Amitabh, and Douglas O. Staiger. 2011. "Expertise, Overuse and Underuse in Healthcare." Unpublished Working Paper.

Chandra, Amitabh, and Douglas O. Staiger. 2020. "Identifying Sources of Inefficiency in Healthcare." *The Quarterly Journal of Economics*, 135(2): 785—-843.

Chen, Irene Y., Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020*a*. "Ethical Machine Learning in Health Care."

Chen, Irene Y., Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. 2020*b*. "Probabilistic Machine Learning for Healthcare." arXiv preprint, arXiv:2009.11087.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2019. "Inference on Causal and Structural Parameters using Many Moment Inequalities." *The Review of Economic Studies*, 86(5): 1867–1900.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen. 2013. "Intersection Bounds: Estimation and Inference." *Econometrica*, 81(2): 667–737.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 134–148.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *The Journal of Finance*, 66(4): 1047–1108.

Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova. 2021. "Characterizing Fairness Over the Set of Good Models Under Selective Labels."

Cowgill, Bo. 2018. "Bias and Productivity in Humans and Machines: Theory and Evidence."

Cox, Gregory, and Xiaoxia Shi. 2020. "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models."

Currie, Janet, and W. Bentley Macleod. 2017. "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians." *Journal of Labor Economics*, 35(1): 1–43.

Currie, Janet, and W. Bentley Macleod. 2020. "Understanding Doctor Decision Making: The Case of Depression Treatment." *Econometrica*, 88(3): 847–878.

Dawes, Robyn M. 1971. "A case study of graduate admissions: Application of three principles of human decision making." *American Psychologist*, 26(2): 180–188.

Dawes, Robyn M. 1979. "The robust beauty of improper linear models in decision making." *American Psychologist*, 34(7): 571–582.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science*, 249(4899): 1668–1674.

De Bondt, Werner F. M., and Richard Thaler. 1985. "Does the Stock Market Overreact?" *Journal of Finance*, 40: 793–805.

de Chaisemartin, Clement. 2017. "Tolerating Defiance? Local Average Treatment Effects Without Monotonicity." *Quantitative Economics*, 8(2): 367–396.

D'Haultfoeuille, Xavier, Christophe Gaillac, and Arnaud Maurel. 2020. "Rationalizing Rational Expectations: Characterization and Tests." arXiv preprint, arXiv:2003.11537.

Dobbie, Will, and Crystal Yang. 2019. "Proposals for Improving the U.S. Pretrial System." The Hamilton Project.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania. 2020. "Measuring Bias in Consumer Lending."

Dobbie, Will, Jacob Goldin, and Crystal Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–240.

Einav, Liran, Mark Jenkins, and Jonathan Levin. 2013. "The impact of credit scoring on consumer lending." *Rand Journal of Economics*, 44(2): 249—-274.

Elliott, Graham, Allan Timmerman, and Ivana Komunjer. 2005. "Estimation and Testing of Forecast Rationality under Flexible Loss." *The Review of Economic Studies*, 72(4): 1107–1125.

Elliott, Graham, Ivana Komunjer, and Allan Timmerman. 2008. "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?" *Journal of the European Economic Association*, 6(1): 122–157.

Erel, Isil, Lea H. Stern, Chenhao Tan, and Michael S. Weisbach. 2019. "Selecting Directors Using Machine Learning." NBER Working Paper Series No. 24435.

Fang, Zheng, Andres Santos, Azeem M. Shaikh, and Alexander Torgovitsky. 2020. "Inference for Large-Scale Linear Systems with Known Coefficients."

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and Removing Disparate Impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Frandsen, Brigham R., Lars J. Lefgren, and Emily C. Leslie. 2019. "Judging Judge Fixed Effects." NBER Working Paper Series No. 25528.

Frankel, Alexander. 2021. "Selecting Applicants." *Econometrica*, 89(2): 615–645.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2018. "Predictably Unequal? The Effects of Machine Learning on Credit Markets."

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *The Quarterly Journal of Economics*, 129(4): 1661–1710.

Gabaix, Xavier. 2019. "Behavioral Inattention." In *Handbook of Behavioral Economics: Applications and Foundations*. Vol. 2, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 261–343. North Holland.

Gelbach, Jonah. 2021. "Testing Economic Models of Discrimination in Criminal Justice."

Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *The Quarterly Journal of Economics*, 125(4): 1399–1433.

Gillis, Talia. 2019. "False Dreams of Algorithmic Fairness: The Case of Credit Pricing."

Grove, W. M., D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson. 2000. "Clinical versus mechanical prediction: A meta-analysis." *Psychological Assessment*, 12(1): 19–30.

Gualdani, Christina, and Shruti Sinha. 2020. "Identification and Inference in Discrete Choice Models with Imperfect Information." arXiv preprint, arXiv:1911.04529.

Gul, Faruk, and Wolfgang Pesendorfer. 2006. "Random Expected Utility." *Econometrica*, 74(1): 121–146.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

Heckman, James J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica*, 42(4): 679–694.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1): 153–161.

Henry, Marc, Romuald Meango, and Ismael Mourifie. 2020. "Revealing Gender-Specific Costs of STEM in an Extended Roy Model of Major Choice."

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2018. "Discretion in Hiring." *The Quarterly Journal of Economics*, 133(2): 765—-800.

Ho, Kate, and Adam M. Rosen. 2017. "Partial Identification in Applied Research: Benefits and Challenges." *Advances in Economics and Econometrics: Eleventh World Congress*, , ed. Bo Honoré, Ariel Pakes, Monika Piazzesi and Larry Samuelson Vol. 2, 307–359. Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81: 945–960.

Hull, Peter. 2021. "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making." NBER Working Paper Series No. 28503.

Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *American Economic Review*, 93(2): 126–132.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62: 467–475.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics*, 26(1): 101–136.

Jansen, Mark, Hieu Nguyen, and Amin Shams. 2021. "Rise of the Machines: The Impact of Automated Underwriting."

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2020*a*. "Simple rules to guide expert classifications." *Journal of the Royal Statistical Society Series A*, 183(3): 771–800.

Jung, Jongbin, Ravi Shroff, Avi Feller, and Sharad Goel. 2020*b*. "Bayesian Sensitivity Analysis for Offline Policy Evaluation." *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 64–70.

Kahneman, Daniel. 2019. "Artificial Intelligence and Behavioral Economics: Comment." In *The Economics of Artificial Intelligence: An Agenda*. , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb, 608–610. University of Chicago Press.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou. 2018. "Interval Estimation of Individual-Level Causal Effects Under Unobserved Confounding." arXiv preprint arXiv:1810.02894.

Kamenica, Emir. 2019. "Bayesian Persuasion and Information Design." *Annual Review of Economics*, 11: 249–272.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review*, 101: 2590–2615.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance*, 34(11): 2767 – 2787.

Kitagawa, Toru. 2020. "The Identification Region of the Potential Outcome Distributions under Instrument Independence." Cemmap Working Paper CWP23/20.

Kitamura, Yuichi, and Jorg Stoye. 2018. "Nonparametric Analysis of Random Utility Models." *Econometrica*, 86(6): 1883–1909.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review: Papers and Proceedings*, 105(5): 491–495.

Kuncel, Nathan R., David M. Klieger, Brian S. Connelly, and Deniz S Ones. 2013. "Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis." *Journal of Applied Psychology*, 98(6): 1060—1072.

Lakkaraju, Himabindu, and Cynthia Rudin. 2017. "Learning Cost-Effective and Interpretable Treatment Regimes." *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54: 166–175.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Li, Danielle, Lindsey Raymond, and Peter Bergman. 2020. "Hiring as Exploration." NBER Working Paper Series No. 27736.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. "Delayed Impact of Fair Machine Learning." *Proceedings of the 35th International Conference on Machine Learning*.

Lu, Jay. 2016. "Random Choice and Private Information." *Econometrica*, 84(6): 1983–2027.

Lu, Jay. 2019. "Bayesian Identification: A Theory for State-Dependent Utilities." *American Economic Review*, 109(9): 3192–3228.

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources*, 24(3): 343–360.

Manski, Charles F. 2017. "Improving Clinical Guidelines and Decisions Under Uncertainty." NBER Working Paper No. 23915.

Marquardt, Kelli. 2021. "Mis(sed) Diagnosis: Physician Decision Making and ADHD."

Martin, Daniel, and Phillip Marx. 2021. "A Robust Test of Prejudice for Discrimination Experiments."

Meehl, Paul E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press.

Molinari, Francesca. 2020. "Microeconometrics with Partial Identification." In *Handbook of Econometrics*. Vol. 7, 355–486.

Mourifie, Ismael, Marc Henry, and Romuald Meango. 2019. "Sharp bounds and testability of a Roy model of STEM major choices." arXiv preprint arXiv:1709.09284.

Mullainathan, Sendhil, and Ziad Obermeyer. 2020. "Who is tested for heart attack and who should be: predicting patient risk and physician error." NBER Working Paper Series, Working Paper No. 26168.

Natenzon, Paulo. 2019. "Random Choice and Learning." *Journal of Political Economy*, 127(1): 419–457.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." 469–481.

Rambachan, Ashesh, and Jonathan Roth. 2020. "An Honest Approach to Parallel Trends."

Rehbeck, John. 2020. "Revealed Bayesian Expected Utility with Limited Data."

Ribers, Michael Allan, and Hannes Ullrich. 2019. "Battling Antibiotic Resistance: Can Machine Learning Improve Prescribing?" arXiv preprint arXiv:1906.03044.

Ribers, Michael Allan, and Hannes Ullrich. 2020. "Machine Predictions and Human Decisions with Variation in Payoffs and Skills."

Rosenbaum, Paul R. 2002. *Observational Studies.* Springer.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika*, 63(3): 581–592.

Russell, Thomas M. 2019. "Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects." *Journal of Business & Economic Statistics*.

Shiller, Robert J. 1981. "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review*, 71: 421–436.

Sims, Christopher A. 2003. "Implications of rational inattention." *Journal of Monetary Economics*, 50(3): 665–690.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review*, 103.

Stevenson, Megan, and Jennifer Doleac. 2019. "Algorithmic Risk Assessment in the Hands of Humans."

Syrgkanis, Vasilis, Elie Tamer, and Juba Ziani. 2018. "Inference on Auctions with Weak Assumptions on Information." arXiv preprint, arXiv:1710.03830.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.

Woodford, Michael. 2013. "Macroeconomic Analysis Without the Rational Expectations Hypothesis." *Annual Review of Economics*, 5(1): 303–346.

Yadlowsky, Steve, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. 2020. "Bounds on the conditional and average treatment effect with unobserved confounding factors." arXiv preprint arXiv:1808.09521.

Yang, Crystal, and Will Dobbie. 2020. "Equal Protection Under Algorithms: A New Statistical and Legal Framework." *Michigan Law Review*, 119(2): 291–396.

A Additional Results for the Expected Utility Maximization Model

In this section of the appendix, I provide additional results for the expected utility maximization model that are mentioned in the main text.

A.1 Expected Utility Maximization Behavior with Inaccurate Beliefs after Dimension Reduction

In this section, I show that the identification result for expected utility maximization behavior with inaccurate beliefs extends to coarsening the excluded characteristics. Let $D_w \colon \mathcal{X} \to \{1, \ldots, d_w\}$ be a function that partitions the observable characteristics X into level sets $\{x \in \mathcal{X} \colon D_w(x) = d\}$. Following Lemma C.3 in the proof of Theorem 4.1, I state a result for the case in which the decision maker faces many choice $\mathcal{C}$ with $|\mathcal{C}| = N_c$, and the researcher only observes the outcome of interest if $C \in \mathcal{C}^y \subseteq \mathcal{C}$ for some known subset $\mathcal{C}^y$.

Proposition A.1. *Suppose the human DM's choices are consistent with expected utility maximization behavior at inaccurate beliefs and some utility function $U \in \mathcal{U}$. Then, for each $w \in \mathcal{W}$, $d \in \{1, \ldots, N_d\}$, $c \in \mathcal{C}^y$ and $c' \neq c$,*

$$\sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c, y^* \mid w, D_w(x) = d) U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c, y^* \mid w, D_w(x) = d) U(c, y^*; w),$$

where

$$Q_{C,Y^*}(c, y^* \mid w, D_w(x) = d) = \left(\sum_{x \colon D_w(x) = d} P_C(c \mid y^*, w, x) \widetilde{P}_{Y^*}(y^* \mid w, x) P(x \mid w) \right) / P(D_w(x) = d \mid w)$$

$$P_C(c \mid y^*, w, x) = \begin{cases} \frac{P_{C,Y^*}(c, y^* \mid w, x)}{\sum_{c \in \mathcal{C} \setminus \mathcal{C}^y} \widetilde{P}_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x) + \sum_{c \in \mathcal{C}^y} P_{C,Y^*}(c, y^* \mid w, x)} & \text{if } c \in \mathcal{C}^y \\ \frac{\widetilde{P}_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x)}{\sum_{c \in \mathcal{C} \setminus \mathcal{C}^y} \widetilde{P}_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x) + \sum_{c \in \mathcal{C}^y} P_{C,Y^*}(c, y^* \mid w, x)} & \text{if } c \in \mathcal{C} \setminus \mathcal{C}^y. \end{cases}$$

Proof. This follows from applying iterated expectations to Lemma C.3. □

Provided that $P_{C,Y^*}(c, y^* \mid w, x) > 0$ for all $(c, y^*) \in \mathcal{C} \times \mathcal{Y}$ and $(w, x) \in \mathcal{W} \times \mathcal{X}$, Proposition A.1 can be recast as checking whether

$$\sum_{y^* \in \mathcal{Y}} \omega(c, y^*; w, d) P_{C,Y^*}(c, y^* \mid w, D_w(x) = d) U(c, y^*; w) \geq$$

$$\sum_{y^* \in \mathcal{Y}} \omega(c, y^*; w, d) P_{C,Y^*}(c, y^* \mid w, D_w(x) = d) U(c, y; w)$$

for non-negative weights $\omega(c, y^*; w, d) \geq 0$ satisfying $\mathbb{E}_P \left[\omega(C, Y^*; W, D_w(X)) \mid W = w, D_w(x) = d \right] = 1$. In a binary screening decision, this result may then be applied to derive bounds on the decision

maker's reweighed utility threshold through

$$P(Y^* = 1 \mid C = 1, W = w, D_w(X) = d) \leq \frac{\omega(0,0;w,d)U(0,0;w)}{\omega(0,0;w,d)U(0,0;w) + \omega(1,1;w,d)U(1,1;w)}$$
$$\frac{\omega(0,0;w,d)U(0,0;w)}{\omega(0,0;w,d)U(0,0;w) + \omega(1,1;w,d)U(1,1;w)} \leq \overline{P}(Y^* = 1 \mid C = 0, W = w, D_w(X) = d).$$

Next, define $M = 1\{C = 0, Y^* = 0\} + 1\{C = 1, Y^* = 1\}$, $\tau(w,d) = \frac{\omega(0,0;w,d)U(0,0;w)}{\omega(0,0;w,d)U(0,0;w)+\omega(1,1;w,d)U(1,1;w)}$. Examining $w \in \mathcal{W}$, $d, d' \in \{1, \ldots, N_d\}$, we arrive at

$$\frac{\frac{Q(C=1,Y^*=1|M=1,w,d)/Q(C=0,Y^*=0|M=1,w,d)}{Q(C=1,Y^*=1|M=1,w,d')/Q(C=0,Y^*=0|M=1,w,d')}}{\frac{P(C=1,Y^*=1|M=1,w,d)/P(C=0,Y^*=0|M=1,w,d)}{P(C=1,Y^*=1|M=1,w,d')/P(C=0,Y^*=0|M=1,w,d')}} = \frac{(1-\tau(w,d))/\tau(w,d)}{(1-\tau(w,d'))/\tau(w,d')}.$$

By examining values in the identified set of reweighted utility thresholds defined on the coarsened characteristic space, bounds may be constructed on a parameter that summarizes the human DM's beliefs about her own "ex-post mistakes." How does the human DM's belief about the relative probability of choosing $C = 0$ and outcome $Y^* = 0$ occurring vs. choosing $C = 1$ and outcome $Y^* = 1$ occurring compare to the true probability. If these bounds lie everywhere below one, then the human DM's beliefs are under-reacting to variation in risk across the cells (w, d) and (w, d'). If these bounds lie everywhere above one, then the human DM's beliefs are over-reacting.

B Additional Results for the Econometric Framework

In this section of the appendix, I provide additional results for the econometric framework presented in Section 3 of the main text.

B.1 Constructing Bounds on the Missing Data through a Quasi-Random Instrument

In this section, I modify Assumption 3.1 to only impose that the instrument be quasi-randomly assigned conditional on some additional characteristics $t \in \mathcal{T}$ with finite support. The joint distribution $(W, X, T, Z, C, Y^*) \sim P$ satisfies

$$Y^* \perp\!\!\!\perp Z \mid W, X, T \tag{4}$$

and $P(W = w, X = x, T = t, Z = z) > 0$ for all $(w, x, t, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{T} \times \mathcal{Z}$.

Under (4), researchers can derive bounds on the unobservable choice-dependent outcome probabilities in a binary screening decision. By iterated expectations,

$$P(Y^* = 1 \mid W = w, X = x, Z = z) =$$
$$\sum_{t \in \mathcal{T}} P(Y^* = 1 \mid W = w, X = x, Z = z, T = t) P(T = t \mid W = w, X = x, Z = z) =$$
$$\sum_{t \in \mathcal{T}} P(Y^* = 1 \mid W = w, X = x, T = t) P(T = t \mid W = w, X = x, Z = z),$$

where the last equality follows by quasi-random assignment. For each value of $t \in \mathcal{T}$ and $z \in \mathcal{Z}$,

$P(Y^* = 1|W = w, X = x, Z = z, T = t)$ is bounded by

$$
\begin{aligned}
&P(Y^* = 1, C = 1 \mid W = w, X = x, Z = z, T = t) \leq \\
&P(Y^* = 1 \mid W = w, X = x, Z = z, T = t) \leq \\
&P(C = 0 \mid W = w, X = x, Z = z, T = t) + P(Y^* = 1, C = 1 \mid W = w, X = x, Z = z, T = t).
\end{aligned}
$$

Therefore, for a given $z \in \mathcal{Z}$, valid lower and upper bounds on $P(Y^* = 1 \mid W = w, X = x, Z = z)$ are given by

$$
\begin{aligned}
&\mathbb{E}\left[P(C = 1, Y^* = 1 \mid W, X, Z = \tilde{z}, T) \mid W = w, X = x, Z = z\right] \leq \\
&P(Y^* = 1 \mid W = w, X = x, Z = z) \leq \\
&\mathbb{E}\left[P(C = 1, Y^* = 1 \mid W, X, Z = \tilde{z}, T) + P(C = 0 \mid W, X, Z = \tilde{z}, T) \mid W = w, X = x, Z = z\right]
\end{aligned}
$$

for any $\tilde{z} \in \mathcal{Z}$. Since $P(C = 1, Y^* = 1 \mid W = w, X = x, Z = z)$ is observed, this naturally implies bounds on $P(C = 0, Y^* = 1 \mid W = w, X = x, Z = z)$, which in turn gives a bound on $P(Y^* = 1 \mid C = 0, W = w, X = x, Z = z)$ since $P(C = 0 \mid W = w, X = x, Z = z)$ is also observed.

B.2 Testing Expected Utility Maximization Behavior in General Screening Decisions

In this section, I extend the econometric framework discussed in Section 3 to general screening decisions. First, I discuss how the researcher may construct bounds on the unobservable choice-dependent outcome probabilities in these settings. Second, I show how testing the revealed preference inequalities in these settings reduces testing many moment inequalities with linear nuisance parameters.

B.2.1 Constructing Bounds with an Instrument in General Screening Decisions

Let $P_{C,Y^*}(\cdot, \cdot \mid w, x, z) \in \Delta(\{0, 1\} \times \mathcal{Y})$ denote the conditional joint distribution of choices and outcomes given $W = w, X = x, Z = z$ and $|\mathcal{Y}| := N_y$. Under Assumption 3.1, this conditional joint distribution is partially identified and the next result provides bounds on this quantity.

Proposition B.1. *Suppose Assumption 3.1 holds and consider any $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$. If $\widetilde{P}_{C,Y^*}(\cdot, \cdot \mid w, x, z) \in \mathcal{H}_P(P_{C,Y^*}(\cdot, \cdot \mid w, x, z))$, then it satisfies*

i. For all $y^* \in \mathcal{Y}$, $\widetilde{P}(C = 1, Y^* = y^* \mid W = w, X = x, Z = z) = P(C = 1, Y = y^* \mid W = w, X = x, Z = z)$;

ii. $\sum_{y^* \in \mathcal{Y}} \widetilde{P}(C = 0, Y^* = y^* \mid W = w, X = x, Z = z) = P(C = 0 \mid W = w, X = x, Z = z)$;

iii. For all $\tilde{z} \in \mathcal{Z}$ with $\tilde{z} \neq z$ and $y^* \in \mathcal{Y}$

$$
\begin{aligned}
&P(C = 1, Y = y^* \mid W = w, X = x, Z = \tilde{z}) \leq \\
&\sum_{c \in \mathcal{C}} \widetilde{P}(C = c, Y^* = y^* \mid W = w, X = x, Z = z) \leq
\end{aligned}
$$

$$P(C = 1, Y = y^* \mid W = w, X = x, Z = \tilde{z}) + P(C = 0 \mid W = w, X = x, Z = \tilde{z})$$

Proof. Consider a particular value $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$. Under Assumption 3.1,

$$P(Y^* = y^* \mid W = w, X = x, Z = z) = P(Y = y^* \mid W = w, X = x, Z = \tilde{z})$$

for all $y^* \in \mathcal{Y}$, $\tilde{z} \neq z$, where $P(Y^* = y^* \mid W = w, X = x, Z = \tilde{z}) = P(C = 1, Y^* = y^* \mid W = w, X = x, Z = \tilde{z}) + P(C = 0, Y^* = y^* \mid W = w, X = x, Z = \tilde{z})$. Therefore, for any $\tilde{z} \neq z$, valid lower and upper bounds on $P(Y^* = y^* \mid W = w, X = x, Z = z)$ are given by

$$P(C = 1, Y = y^* \mid W = w, X = x, Z = \tilde{z}) \leq$$

$$P(Y^* = y^* \mid W = w, X = x, Z = z) \leq$$

$$P(C = 1, Y = y^* \mid W = w, X = x, Z = \tilde{z}) + P(C = 0 \mid W = w, X = x, Z = \tilde{z}).$$

The result follows by noting that $P(C = 1, Y^* = y^* \mid W = w, X = x, z = z)$ is observed for all $y^* \in \mathcal{Y}$ and $P(C = 0 \mid W = w, X = x, Z = z)$ is also observed. □

The researcher may then construct bounds on the unobservable choice dependent outcome probabilities since the decision maker's conditional choice probabilities are observed.

These simple bounds are non-sharp in general, but sharp bounds can be derived using Arstein's theorem, which follows results in Russell (2019); Kitagawa (2020) who characterize the identified set of the joint distribution of potential outcomes in treatment assignment problems. I will drop the conditioning on $(w, x) \in \mathcal{W} \times \mathcal{X}$ for conciseness and this argument applies conditionally on each $(w, x) \in \mathcal{W} \times \mathcal{X}$. The screening decision setting establishes the model correspondence $G \colon \mathcal{Y} \to \mathcal{Z} \times \mathcal{C} \times (\mathcal{Y} \cup \{0\})$, where $G(y^*) = \{(z, c, y) \colon y = y^* 1\{c = 1\}\}$. The reverse correspondence is given by $G^{-1}(z, c, y) = \{y^* \colon y = y^* 1\{c = 1\}\}$. The observable joint distribution $(Z, C, Y) \sim P$ characterizes a random set $G^{-1}(Z, C, Y)$ via the generalized likelihood $T(A \mid Z = z) = P\left((C, Y) \colon G^{-1}(z, C, Y) \cap A \neq \emptyset\right)$ for all $A \in 2^{\mathcal{Y}}$. Artstein's Theorem implies that there exists a random variable Y^* that rationalizes the observed data through the model correspondence G if and only if there exists some $Y^* \sim \widetilde{P}$ satisfying

$$\widetilde{P}(A) \leq T(A \mid Z = z) \text{ for all } A \in 2^{\mathcal{Y}} \text{ and } z \in \mathcal{Z}. \tag{5}$$

Let $\widetilde{\mathcal{P}}_{Y^*}(\cdot \mid z)$ be the set of distributions on $\mathcal{Y}$ that satisfy these inequalities at a given $z \in \mathcal{Z}$. A sharp characterization of identified set for the marginal distribution of Y^* is then given by $\mathcal{H}_P(P_{Y^*}(\cdot)) = \bigcap_{z \in \mathcal{Z}} \widetilde{\mathcal{P}}_{Y^*}(\cdot \mid z)$. This delivers sharp bounds on $\widetilde{P}(C = 0, Y^* = y^*)$ since $\widetilde{P}(C = 1, Y^* = y^*)$ is observed. In general, Equation (5) implies $2^{N_y} - 2$ inequalities associated with each choice of the instrument. Therefore, the number of inequalities in the characterization of the sharp identified set will grow exponentially in the number of support points of the outcome.

To illustrate these sharp inequalities, consider a binary screening decision. Artstein's Theorem gives for each $z \in \mathcal{Z}$

$$\widetilde{P}(Y^* = 0) \leq P(C = 0 \mid Z = z) + P(C = 1, Y = 0 \mid Z = z)$$
$$\widetilde{P}(Y^* = 1) \leq P(C = 0 \mid Z = z) + P(C = 1, Y = 1 \mid Z = z).$$

Since $\widetilde{P}(Y^* = 0) + \widetilde{P}(Y^* = 1) = 1$, these inequalities maybe reduced to simply requiring for each $z \in \mathcal{Z}$

$$P(C = 1, Y = 1 \mid Z = z) \leq \widetilde{P}(Y^* = 1) \leq P(C = 0 \mid Z = z) + P(C = 1, Y = 1 \mid Z = z).$$

This corresponds to the inequalities stated in Proposition 3.1 of the main text.

B.2.2 Testing for Prediction Mistakes Reduces to Testing Many Moment Inequalities with Linear Nuisance Parameters

Suppose the researcher wishes to test whether the decision maker's choices are consistent with expected utility maximization behavior at some utility function $U \in \mathcal{U}$ in a general screening decision, where recall that $\mathcal{U}$ is the set of feasible utility functions specified by the researcher. Denote this null hypothesis by $H_0(\mathcal{U})$ As a stepping stone, I provide a reduction to show how the researcher may test whether the decision maker's choices are consistent with expected utility maximization behavior at a particular utility function $U \in \mathcal{U}$. Denote this particular null hypothesis as $H_0(U)$. As discussed in Bugni, Canay and Shi (2015), the researcher may construct a conservative test of $H_0(\mathcal{U})$ by constructing a confidence interval for the identified set of utility functions through test inversion on $H_0(U)$ and checking whether this confidence interval is empty.

With this in mind, consider a fixed utility function $U \in \mathcal{U}$ and suppose the researcher constructs bounds on the unobservable choice-dependent outcome probabilities using a randomly assigned instrument. Testing $H_0(U)$ is equivalent to testing a possibly high-dimensional set of moment inequalities with linear nuisance parameters.

Proposition B.2. *Suppose Assumption 3.1 holds. Letting $N_y := |\mathcal{Y}|$, the decision maker's choices at $z \in \mathcal{Z}$ are consistent with expected utility maximization behavior at utility function $U : \mathcal{W} \times \{0, 1\} \times \mathcal{Y} \to \mathbb{R}$ if there exists $\tilde{\delta} \in \mathbb{R}^{d_w d_x (N_y - 1)}$ satisfying*

$$\tilde{A}(U)_{(\cdot, 1:m)} \mu(P) + \tilde{A}(U)_{(\cdot, -(1:m))} \tilde{\delta} \leq b,$$

where $\tilde{A}(U)$ is a matrix of known constants that depend on the specified utility function U, b is a vector of known constants, $\mu(P)$ is a $m := 2\left(d_w d_x N_y + d_w d_x N_y (N_z - 1)\right)$ dimensional vector that collects together the observable moments and bounds based on the instrument.[11]

Proof. To prove this result for the binary choice $\mathcal{C} = \{0, 1\}$, I will instead prove a more general result in which the decision maker may faces many possible choice as in the proof of Theorem 2.1 with $|\mathcal{C}| = N_c$.

Applying the non-sharp bounds in Proposition B.1, I begin by restating Lemma C.1 in terms

[11] For a matrix B, the notation $B_{(\cdot, 1:m)}$ denotes the submatrix containing the first m columns of B and $B_{(\cdot, -1:m)}$ denotes the submatrix containing all columsn except the first m of B.

of the nuisance parameters $\widetilde{P}_{C,Y^*}(\cdot \mid w,x,z) \in \Delta(\mathcal{C}\times\mathcal{Y})$ with entries

$$\widetilde{P}_{C,Y^*}(\cdot \mid w,x,z) = \begin{pmatrix} P_{C,Y^*}(c_1, y_1 \mid w,x,z) \\ \vdots \\ P_{C,Y^*}(c_1, y_{N_y} \mid w,x,z) \\ \vdots \\ P_{C,Y^*}(c_{N_c}, y_1 \mid w,x,z) \\ \vdots \\ P_{C,Y^*}(c_{N_c}, y_{N_y} \mid w,x,z) \end{pmatrix}.$$

Lemma B.1. *Suppose Assumption 3.1 holds. The human DM's choices at $z \in \mathcal{Z}$ are consistent with expected utility maximization behavior at utility function U if for all $(w,x) \in \mathcal{W}\times\mathcal{X}$ there exists $\widetilde{P}_{C,Y^*}(\cdot \mid w,x,z) \in \Delta(\mathcal{C}\times\mathcal{Y})$ satisfying*

i. For all $c \in \mathcal{C}$, $\tilde{c} \neq c$,

$$\sum_{y\in\mathcal{Y}} \widetilde{P}_{C,Y^*}(c,y \mid w,x,z)U(c,y;w,z) \geq \sum_{y\in\mathcal{Y}} \widetilde{P}_{C,Y^*}(c,y \mid w,x,z)U(\tilde{c},y;w,z).$$

ii. For all $c \in \mathcal{C}^y$ and $y \in \mathcal{Y}$, $\widetilde{P}_{C,Y^*}(c,y \mid w,x,z) = P_{C,Y^*}(c,y \mid w,x,z)$.
For all $c \in \mathcal{C}\setminus\mathcal{C}^y$, $\sum_{y\in\mathcal{Y}} \widetilde{P}_{C,Y^*}(c,y \mid w,x,z) = P_{C,Y^*}(c \mid w,x,z)$.

iii. For all $y \in \mathcal{Y}$ and $\tilde{z} \in \mathcal{Z}$,

$$P_{C,Y^*}(C\in\mathcal{C}^y, Y=y \mid W=w, X=x, Z=\tilde{z}) \leq$$

$$\sum_{c\in\mathcal{C}} \widetilde{P}_{C,Y^*}(C=c, Y=y \mid W=w, X=x, Z=z) \leq$$

$$P_{C,Y^*}(C\in\mathcal{C}^y, Y=y \mid W=w, X=x, Z=\tilde{z}) + P_C(C\in\mathcal{C}\setminus\mathcal{C}^y \mid W=w, X=x, Z=\tilde{z}).$$

For each $c \in \mathcal{C}$ and $\tilde{c} \neq c$, define the $1\times N_y$ dimensional row vector $A^{c,\tilde{c}}_{w,x,z}(U)$ as

$$A^{c,\tilde{c}}_{w,x,z}(U) = \left(U(\tilde{c},y_1;w,z) - U(c,y_1;w,z), \ldots, U(\tilde{c},y_{N_y};w,z) - U(c,y_{N_y};w,z)\right).$$

For each $c \in \mathcal{C}$ define the matrix $A^{c}_{w,x,z}(U)$ to be the $(N_c-1)\times N_y$ matrix whose rows are equal to $A^{c,\tilde{c}}_{w,x,z}(U)$. For each $(w,x) \in \mathcal{W}\times\mathcal{X}$, define the $N_c(N_c-1)\times N_cN_y$ dimensional block diagonal matrix $A_{w,x,z}(U)$ as

$$A_{w,x,z}(U) = \begin{pmatrix} A^{c_1}_{w,x,z}(U) & & & \\ & A^{c_2}_{w,x,z}(U) & & \\ & & \ddots & \\ & & & A^{c_{N_c}}_{w,x,z}(U) \end{pmatrix}. \tag{6}$$

Define the $d_w \cdot d_x \cdot N_c(N_c-1) \times d_w \cdot d_x \cdot N_c N_y$ dimensional block diagonal matrix $A_z(U)$ as

$$A_z(U) = \begin{pmatrix} A_{w_1,x_1,z}(U) & & & \\ & A_{w_1,x_2,z}(U) & & \\ & & \ddots & \\ & & & A_{w_{d_w},x_{d_x},z}(U) \end{pmatrix}. \tag{7}$$

Letting $\widetilde{P}_{C,Y^*}(\cdot \mid z) = (P_{C,Y^*}(\cdot \mid w_1, x_1, z), \ldots, P_{C,Y^*}(\cdot \mid w_{d_w}, x_{d_x}, z))$, the revealed preference constraints in (i) of Lemma B.1 can be rewritten as

$$A_z(U)\widetilde{P}_{C,Y^*}(\cdot \mid z) \leq 0,$$

where $\widetilde{P}_{C,Y^*}(\cdot \mid z)$ is a $(d_w \cdot d_x \cdot N_c \cdot N_y) \times 1$ dimensional vector.

We may construct a $(d_w \cdot d_x \cdot |\mathcal{C}^y| \cdot N_y) \times (d_w \cdot d_x \cdot N_c \cdot N_y)$ dimensional matrix $B^{\mathcal{C}^y}_{z,eq}$ and a $(d_w \cdot d_x \cdot (N_c - |\mathcal{C}^y|)) \times (d_w \cdot d_x \cdot N_c \cdot N_y)$ dimensional matrix $B^{\mathcal{C}-\mathcal{C}^y}_{z,eq}$ that forms the data consistency conditions in (ii) of Lemma B.1. For each $\tilde{z} \in \mathcal{Z}$, we may construct the $(d_w \cdot d_x \cdot N_y) \times (d_w \cdot d_x \cdot N_c \cdot N_y)$ dimensional matrices $\underline{B}_{z,\tilde{z}}$, $\overline{B}_{z,\tilde{z}}$ that forms the lower and upper bounds respectively in condition (iii) of Lemma B.1. Collect these together into $\underline{B}_z, \overline{B}_z$. Finally, we define the $(d_w \cdot d_x) \times (d_w \cdot d_x \cdot N_c \cdot N_y)$ dimensional matrix $D_{z,eq}$ that imposes that each $\widetilde{P}_{C,Y^*}(\cdot \mid w, x, z)$ sum to one, and the $(d_w \cdot d_x \cdot N_c \cdot N_y) \times (d_w \cdot d_x \cdot N_c \cdot N_y)$ dimensional matrix $D_{z,+}$ that imposes that each element $\widetilde{P}_{C,Y^*}(\cdot \mid w, x, z)$ is non-negative.

Finally, we introduce non-negative slack parameters associated with each inequality constraint in conditions (ii)-(iii). Putting this together, we may then write (ii)-(iii) in Lemma B.1 as

$$\underbrace{\begin{pmatrix} B^{\mathcal{C}^y}_{z,eq} & 0 & 0 \\ B^{\mathcal{C}\setminus\mathcal{C}^y}_{z,eq} & 0 & 0 \\ \underline{B}_z & 1 & 0 \\ \overline{B}_z & 0 & 1 \end{pmatrix}}_{:=B_z} \underbrace{\begin{pmatrix} \widetilde{P}_{C,Y^*}(\cdot \mid z) \\ \underline{s}_z \\ \overline{s}_z \end{pmatrix}}_{:=\delta} = \mu(P)$$

where $\mu(P)$ is the vector that collects together the observable data and the bounds and vectors of ones/zeros associated with these constraints, and δ is a $(d_w \cdot d_x \cdot N_c \cdot N_y) + 2(d_w \cdot d_x \cdot N_y \cdot (N_z - 1))$ dimensional vector. By further introducing the slack variables into the revealed preference inequalities, we observe that Lemma B.1 implies that the human DM's choices at $z \in \mathcal{Z}$ are consistent with expected utility maximization behavior at utility function U if and only if there exists vector δ satisfying

$$\underbrace{\begin{pmatrix} A_z(U) & 0 & 0 \\ D_{z,eq} & 0 & 0 \\ -D_{z,eq} & 0 & 0 \\ D_{z,+} & 0 & 0 \\ 0 & D_{z,+} & 0 \\ 0 & 0 & D_{z,+} \end{pmatrix}}_{:=\tilde{A}_z(U)} \delta \leq \underbrace{\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{:=b} \text{ and } B_z\delta = \mu(P).$$

The matrix B_z has full row rank and

$$\text{nrow}(B_z) = (d_w \cdot d_x \cdot |\mathcal{C}^y| \cdot N_y) + (d_w \cdot d_x \cdot (N_c - |\mathcal{C}^y|)) + 2(d_w \cdot d_x \cdot N_y \cdot (N_z - 1)),$$

$$\text{ncol}(B_z) = d_w \cdot d_x \cdot N_c \cdot N_y + 2(d_w \cdot d_x \cdot N_y \cdot (N_z - 1)).$$

We observe that $\text{nrow}(B_z) \leq \text{ncol}(B_z)$ since $0 \leq (N_c - |\mathcal{C}^y|)(N_y - 1)$. Therefore, we may define $H_z = \begin{pmatrix} B_z \\ \Gamma_z \end{pmatrix}$ to be the full-rank, square matrix that pads B_z with linearly independent rows. Then,

$$\tilde{A}_z(U)\delta = \tilde{A}_z(U)H_z^{-1}H_z\delta = \tilde{A}_z(U)H_z^{-1}\begin{pmatrix} \mu(P) \\ \tilde{\delta} \end{pmatrix},$$

where $\tilde{\delta} := \Gamma_z\delta$ is a $d_w \cdot d_x \cdot (N_c - |\mathcal{C}^y|)(N_y - 1)$ dimensional vector. This completes the proof with the slight abuse of notation by also defining $\tilde{A}_z(U)H_z^{-1}$ to be $\tilde{A}(U)$. □

As shown in the proof, the number of moment inequalities depends on the number of outcomes, characteristics, and instrument values. This will typically be quite large in empirical applications, and so researchers that wish to directly test the revealed preference inequalities over the characteristics will need to resort to testing procedures for moment inequalities with nuisance parameters that are valid in high dimensional settings such as those developed in Belloni, Bugni and Chernozhukov (2018).

Proposition B.2 showed that testing whether the human DM's choices are consistent with expected utility maximization behavior at a candidate utility function may be reduced to testing a many moment inequalities with linear nuisance parameters. This same testing problem may be also reduced to testing whether there exists a non-negative solution to a large, linear system of equations, which was recently studied in Kitamura and Stoye (2018) and Fang et al. (2020).

For each $w \in \mathcal{W}$, define $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ to be some function that partitions the support of the characteristics $x \in \mathcal{X}$ into the level sets $\{x \colon D_w(x) = d\}$. Through an application of iterated expectations, if the decision maker's choices are consistent with expected utility maximization behavior at some utility function U, then her choices must satisfy *implied* revealed preference inequalities.

Corollary B.1. *Suppose the decision maker's choices are consistent with expected utility maximization behavior at some utility function U. Then, for all $w \in \mathcal{W}$ and $d \in \{1, \ldots, N_d\}$,*

$$\mathbb{E}_{Q^*}\left[U(c, Y^*; W) \mid C = c, W = w, D_w(X) = d\right] \geq \mathbb{E}_{Q^*}\left[U(c', Y^*; W) \mid C = c, W = w, D_w(X) = d\right],$$

for all $c \in \{0, 1\}$, $(w, d) \in \mathcal{W} \times \{1, \ldots, N_d\}$ with $P(C = c \mid W = w, D_w(X) = d) > 0$ and $c' \neq c$, where $\mathbb{E}_{Q^*}[\cdot]$ is the expectation under Q^* and for some $\widetilde{P}^y_{c,w,x} \in \mathcal{B}^y_{c,w,x}$ for all $c \in \mathcal{C} \setminus \mathcal{C}^y$, $(w, x) \in \mathcal{W} \times \mathcal{X}$

$$Q^*(W = w, D_w(X) = d, C = 1, Y^* = y^*) = \sum_{x \colon D_w(x) = d} P(W = w, X = x, C = 1, Y^* = y^*),$$

$$Q^*(W = w, D_w(X) = d, C = 0, Y^* = y^*) =$$

$$\sum_{x\colon D_w(x)=d} \widetilde{P}(Y^* = y^* \mid C = 0, W = w, X = x) P(C = 0, W = w, X = x).$$

In general screening decisions, researchers may test the substantially smaller number of moment inequalities with linear nuisance parameters. This is useful as recent work develops computationally tractable and power inference procedures for lower-dimensional moment inequality with linear nuisance parameters such as Andrews, Roth and Pakes (2019), Cox and Shi (2020) and Rambachan and Roth (2020).

C Proofs of Main Results

Proof of Theorem 2.1

To prove this result for a binary choice $\mathcal{C} = \{0, 1\}$, I prove a more general result in which the decision maker faces many possible choices $\mathcal{C}$ with $|\mathcal{C}| := N_c$. The researcher only observes the outcome of interest if $C \in \mathcal{C}^y \subseteq \mathcal{C}$ for some known subset $\mathcal{C}^y$. As in the main text, let $P_{Y^*}(y^* \mid c, w, x) := P(Y^* = y^* \mid C = c, W = w, X = x)$ and let $P(\cdot \mid c, w, x) \in \Delta(\mathcal{Y})$ denote the vector of choice dependent outcome probabilities given $C = c$, $W = w$, $X = x$. The researcher places bounds on the missing choice-dependent outcome probabilities of the form: for each $c \in \mathcal{C} \setminus \mathcal{C}^y$ and $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists a known subset $\mathcal{B}_{c,w,x} \subseteq \Delta(\mathcal{Y})$ such that $P_{Y^*}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$. The definition of expected utility maximization behavior (Definition 3) extends directly this setting.

I prove the following Lemma, and then show that it implies the analogue of Theorem 2.1 in the multiple choice case.

Lemma C.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ that satisfies for all $(w, x) \in \mathcal{W} \times \mathcal{X}$,*

i. For all $c \in \mathcal{C}^y$ and $c' \neq c$,

$$\sum_{y^* \in \mathcal{Y}} P_{C,Y^*}(c, y^* \mid w, x) U(c, y^*; w) \geq \sum_{y \in \mathcal{Y}} P_{C,Y^*}(c, y^* \mid w, x) U(c', y^*; w).$$

ii. For all $c \in \mathcal{C} \setminus \mathcal{C}^y$, there exists $\widetilde{P}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$ such that

$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x) U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} \widetilde{P}_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x) U(c', y^*; w,).$$

for all $c' \neq c$.

Proof of Lemma C.1: Necessity Suppose that the decision maker's choices are consistent with expected utility maximization behavior at some utility function U and joint distribution $(W, X, V, C, Y^*) \sim Q$.

First, I show that if the decision maker's choices are consistent with expected utility maximization behavior at some utility function U, joint distribution $(W, X, V, C, Y^*) \sim Q$ and private information with support $\mathcal{V}$, then her choices are also consistent with expected utility maximization behavior at some finite support private information. I provide the detailed construction for

the binary choice $\mathcal{C} = \{0, 1\}$ and the same construction applies to the multiple choice case. For $\mathcal{C} = \{0, 1\}$, partition the original signal space $\mathcal{V}$ into the subsets $\mathcal{V}_{\{0\}}, \mathcal{V}_{\{1\}}, \mathcal{V}_{\{0,1\}}$, which collect together the signals $v \in \mathcal{V}$ at which the decision maker strictly prefers $C = 0$, strictly prefers $C = 1$ and is indifferent between $C = 0, C = 1$ respectively. Define the finite support signal space $\widetilde{\mathcal{V}} = \{v_{\{0\}}, v_{\{1\}}, v_{\{0,1\}}\}$ and the finite support private information $\widetilde{V} \in \widetilde{\mathcal{V}}$ as

$$\widetilde{Q}(\widetilde{V} = v_{\{0\}} \mid Y^* = y^*, W = w, X = x) = Q(V \in \mathcal{V}_{\{0\}} \mid Y^* = y^*, W = w, X = x)$$
$$\widetilde{Q}(\widetilde{V} = v_{\{1\}} \mid Y^* = y^*, W = w, X = x) = Q(V \in \mathcal{V}_{\{1\}} \mid Y^* = y^*, W = w, X = x)$$
$$\widetilde{Q}(\widetilde{V} = v_{\{0,1\}} \mid Y^* = y^*, W = w, X = x) = Q(V \in \mathcal{V}_{\{0,1\}} \mid Y^* = y^*, W = w, X = x)$$

Define $\widetilde{Q}(C = 0 \mid \widetilde{V} = v_{\{0\}}, W = w, X = x) = 1$, $\widetilde{Q}(C = 1 \mid \widetilde{V} = v_{\{1\}}, W = w, X = x) = 1$ and

$$\widetilde{Q}(C = 1 \mid \widetilde{V} = v_{\{0,1\}}, W = w, X = x) = \frac{Q(C = 1, V \in \mathcal{V}_{\{0,1\}} \mid W = w, X = x)}{Q(V \in \mathcal{V}_{\{0,1\}} \mid W = w, X = x)}.$$

Then, define the finite support expected utility representation for the decision maker by the utility function U and the random vector $(W, X, \widetilde{V}, C, Y^*) \sim \widetilde{Q}$, where $\widetilde{Q}(w, x, v, c, y^*) = Q(w, x, y^*)\widetilde{Q}(\widetilde{v} \mid w, x, y^*)\widetilde{Q}(c \mid w, x, \widetilde{v})$. The information set and expected utility maximization conditions are satisfied by construction. Data consistency is satisfied since it is satisfied at the original private information $V \in \mathcal{V}$. To see this, notice that for all $(w, x, y^*) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$

$$\begin{aligned}
&P(C = 1, Y^* = y^* \mid W = w, X = x) = \\
&Q(C = 1, V = \mathcal{V}, Y^* = y^* \mid W = w, X = x) = \\
&Q(C = 1, V \in \mathcal{V}_{\{1\}}, Y^* = y^* \mid W = w, X = x) + Q(C = 1, V \in \mathcal{V}_{\{0,1\}}, Y^* = y^* \mid W = w, X = x) = \\
&\widetilde{Q}(C = 1, \widetilde{V} = v_{C=1}, Y^* = y^* \mid W = w, X = x) + \widetilde{Q}(C = 1, \widetilde{V} = v_{C=e}, Y^* = y^* \mid W = w, X = x) = \\
&\sum_{\widetilde{v} \in \widetilde{\mathcal{V}}} \widetilde{Q}(C = 1, \widetilde{V} = \widetilde{v}, Y^* = y^* \mid W = w, X = x) = \widetilde{Q}(C = 1, Y^* = y^* \mid W = w, X = x).
\end{aligned}$$

The same argument applies to $P(C = 0, Y^* = y^* \mid W = w, X = x)$. In the multiple choice case with $|\mathcal{C}| = N_c$, the support $\mathcal{V}$ must be partitioned into subsets associated with each possible subset of the choices, $2^{\mathcal{C}}$, based on all possible combinations of strict preferences for particular choices and indifferences between multiple choices. The finite support private information is then defined to take on 2^{N_c} values. Therefore, for the remainder of the necessity proof, it is without loss of generality to assume the private information $V \in \mathcal{V}$ has finite support.

I next show that if there exists an expected utility representation for the decision maker's choices, then the stated inequalities in Lemma C.1 are satisfied by adapting the necessity argument given the "no-improving action switches inequalities" in Caplin and Martin (2015). Suppose that the decision maker's choices are consistent with expected utility maximization behavior at some utility function U and joint distribution $(W, X, V, C, Y^*) \sim Q$. Then, for each $c \in \mathcal{C}$, $(w, x, v) \in$

$\mathcal{W} \times \mathcal{X} \times \mathcal{V}$

$$Q_C(c \mid w,x,v)\left(\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid w,x,v)U(c,y^*;w)\right) \geq Q_C(c \mid w,x,v)\left(\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid w,x,v)U(c',y^*;w)\right)$$

holds for all $c \neq c'$. If $Q_C(c \mid w,x,v) = 0$, this holds trivially. If $Q_C(c \mid w,x,v) > 0$, this holds through the expected utility maximization condition. Multiply both sides by $Q_V(v \mid w,x)$ to arrive at

$$Q_C(c \mid w,x,v)Q_V(v \mid w,x)\left(\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid w,x,v)U(c,y^*;w)\right) \geq$$

$$Q_C(c \mid w,x,v)Q_V(v \mid w,x)\left(\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid w,x,v)U(c',y^*;w)\right).$$

Next, use information set to write $Q_{C,Y^*}(c,y^* \mid w,x,v) = Q_{Y^*}(y^* \mid w,x,v)Q_C(c \mid w,x,v)$ and arrive at

$$Q_V(v \mid w,x)\left(\sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c,y^* \mid w,x,v)U(c,y;w,x)\right) \geq$$

$$Q_V(v \mid w,x)\left(\sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c,y^* \mid w,x,v)U(c',y^*;w)\right).$$

Finally, we use $Q_{C,Y^*}(c,y^*,v \mid w,x) = Q_{C,Y^*}(c,y^* \mid w,x,v)Q_V(v \mid w,x)$ and then further sum over $v \in \mathcal{V}$ to arrive at

$$\sum_{y^* \in \mathcal{Y}}\left(\sum_{v \in \mathcal{V}} Q_{V,C,Y^*}(v,c,y^* \mid w,x)\right)U(c,y^*;w) \geq \sum_{y^* \in \mathcal{Y}}\left(\sum_{v \in \mathcal{V}} Q_{V,C,Y^*}(v,c,y^* \mid w,x)\right)U(c',y^*;w)$$

$$\sum_{y \in \mathcal{Y}} Q_{C,Y^*}(c,y^* \mid w,x)U(c,y^*;w) \geq \sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c,y^*,w,x)U(c',y^*;w)$$

The inequalities in Lemma C.1 then follow from an application of data consistency.

Proof of Lemma C.1: Sufficiency To establish sufficiency, I show that if the conditions in Lemma C.1 holds, then private information $v \in \mathcal{V}$ can be constructed that recommends choices $c \in \mathcal{C}$ and an expected utility maximizer would find it optimal to follow these recommendations as in the sufficiency argument in Caplin and Martin (2015) for the "no-improving action switches" inequalities.

Towards this, suppose that the conditions in Lemma C.1 are satisfied at some $\widetilde{P}_{Y^*}(\cdot \mid c,w,x) \in \mathcal{B}_{c,w,x}$ for all $c \in \mathcal{C} \backslash \mathcal{C}^y$, $(w,x) \in \mathcal{W} \times \mathcal{X}$. As notation, let $v \in \mathcal{V} := \{1,\ldots,2^{N_c}\}$ index all possible subsets in the power set $2^{\mathcal{C}}$.

For each $(w,x) \in \mathcal{W} \times \mathcal{X}$, define $\mathcal{C}_{w,x} := \{c\colon P_C(c \mid w,x) > 0\} \subseteq \mathcal{C}$ to be the set of choices selected with positive probability, and partition $\mathcal{C}_{w,x}$ into subsets that have identical choice-dependent outcome probabilities. There are $\bar{V}_{w,x} \leq |\mathcal{C}_{w,x}|$ such subsets. Each subset of this partition of $\mathcal{C}_{w,x}$ is a subset in the power set $2^{\mathcal{C}}$, and so I may associate each subset in this par-

tition with its associated index $v \in \mathcal{V}$. Denote the choice-dependent outcome probability associated with the subset labelled v by $P_{Y^*}(\cdot \mid v, w, x) \in \Delta(\mathcal{Y})$. Finally, define $Q_{Y^*}(y^* \mid w, x) = \sum_{c \in \mathcal{C}^y} P_{C,Y^*}(c, y^* \mid w, x) + \sum_{c \in \mathcal{C} \setminus \mathcal{C}^y} \widetilde{P}_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x)$.

Define the random variable $V \in \mathcal{V}$ according to

$$Q_V(v \mid w, x) = \sum_{c:\, P_{Y^*}(\cdot|c,w,x) = P_{Y^*}(\cdot|v,w,x)} P_C(c \mid w, x) \text{ if } v \in \mathcal{V}_{w,x},$$

$$Q_V(v \mid y^*, w, x) = \begin{cases} \frac{P_{Y^*}(y^*|v,w,x) Q(V=v|w,x)}{Q_{Y^*}(y^*|w,x)} \text{ if } v \in \mathcal{V}_{w,x} \text{ and } Q_{Y^*}(y^* \mid w, x) > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Next, define the random variable $C \in \mathcal{C}$ according to

$$Q_C(c \mid v, w, x) = \begin{cases} P_C(c \mid w, x) \Big/ \left(\sum_{\tilde{c}:\, P_{Y^*}(\cdot|\tilde{c},w,x) = P_{Y^*}(\cdot|v,w,x)} P_C(\tilde{c} \mid w, x) \right) \text{ if } v \in \mathcal{V}_{w,x} \text{ and } P_{Y^*}(\cdot \mid c, w, x) = P_{Y^*}(\cdot \mid v, w, x) \\ 0 \text{ otherwise.} \end{cases}$$

Together, this defines the random vector $(W, X, Y^*, V, C) \sim Q$, whose joint distribution is defined as

$$Q(w, x, y^*, v, c) = P_{W,X}(w, x) Q_{Y^*}(y^* \mid w, x) Q_V(V = v \mid y^*, w, x) Q_C(c \mid v, w, x).$$

We now check that this construction satisfies information set, expected utility maximization and data consistency. First, information set is satisfied since $Q_{C,Y^*}(c, y^* \mid w, x, v) = Q_{Y^*}(y^* \mid w, x, v) Q_C(c \mid w, x, v)$ by construction. Next, for any $(w, x) \in \mathcal{W} \times \mathcal{X}$ and $c \in \mathcal{C}_{w,x}$, define $v_{c,w,x} \in \mathcal{V}_{w,x}$ to be the label satisfying $P_{Y^*}(\cdot \mid c, w, x) = P_{Y^*}(\cdot \mid v, w, x)$. For $P_{C,Y}(c, y^* \mid w, x) > 0$, observe that

$$P_{C,Y^*}(c, y^* \mid w, x) =$$
$$P_{Y^*}(y^* \mid c, w, x) P_C(c \mid w, x) =$$
$$Q_{Y^*}(y^* \mid w, x) \frac{Q_{Y^*}(y^* \mid v_{c,w,x}, w, x) \sum_{\tilde{c}:\, P_{Y^*}(\cdot|\tilde{c},w,x) = P_{Y^*}(\cdot|\tilde{c},w,x)} P_C(\tilde{c} \mid w, x)}{Q_{Y^*}(y^* \mid w, x)} \frac{P_C(c \mid w, x)}{\sum_{\tilde{c}:\, P_{Y^*}(\cdot|\tilde{c},w,x) = P_{Y^*}(\cdot|\tilde{c},w,x)} P_C(\tilde{c} \mid w, x)} =$$
$$Q_{Y^*}(y^* \mid w, x) Q_V(v_{c,w,x} \mid y^*, w, x) Q_C(c \mid v_{c,w,x}, w, x) =$$
$$\sum_{v \in \mathcal{V}} Q_{Y^*}(y^* \mid w, x) Q_V(v \mid y^*, w, x) Q_C(c \mid v, w, x) = \sum_{v \in \mathcal{V}} Q_{V,C,Y^*}(v, c, y^* \mid w, x) = Q_{C,Y^*}(c, y^* \mid w, x).$$

Moreover, whenever $P_{C,Y^*}(c, y^* \mid w, x) = 0$, $Q_{Y^*}(y^* \mid v_{c,w,x}, w, x) Q_C(c \mid v_{c,w,x}, w, x) = 0$. Therefore, data consistency holds. Finally, by construction, for $Q_C(C = c \mid V = v_{c,w,x}, W =$

$w, X = x) > 0$,

$$Q(Y^* = y^* \mid V = v_{c,w,x}, W = w, X = x) = \\ \frac{Q(V = v_{c,w,x} \mid Y^* = y^*, W = w, X = x) Q(Y^* = y^* \mid W = w, X = x)}{Q(V = v_{c,w,x} \mid W = w, X = x)} = \\ P(Y^* = y^* \mid C = c, W = w, X = x).$$

Therefore, expected utility maximization is satisfied since the inequalities in Lemma C.1 were assumed to hold and data consistency holds.

Lemma C.1 implies Theorem 2.1: Define the joint distribution Q as

$$Q(w, x, c, y^*) = \begin{cases} P(w, x, c, y^*) \text{ if } c \in \mathcal{C}^y, \\ \widetilde{P}_{Y^*}(y^* \mid c, w, x) P(c, w, x) \text{ if } c \in \mathcal{C} \setminus \mathcal{C}^y. \end{cases}$$

Then, rewrite conditions (i)-(ii) in Lemma C.1 as: for all $c \in \mathcal{C}$ and $c' \neq c$,

$$\sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c, y^* \mid w, x) U(c, y; w) \geq \sum_{y \in \mathcal{Y}} Q_{C,Y^*}(c, y^* \mid w, x) U(c', y; w).$$

Notice that if $P_C(c \mid w, x) = 0$, then $Q_{C,Y^*}(c, y^* \mid w, x) = 0$. Therefore, the inequalities involving $c \in \mathcal{C}$ with $P_C(c \mid w, x) = 0$ are satisfied. Next, inequalities involving $c \in \mathcal{C}$ with $P_C(c \mid w, x) > 0$ can be equivalently rewritten as

$$\sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(y \mid c, w, x) U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(y^* \mid c, w, x) U(c', y; w).$$

The statement of Theorem 2.1 follows by noticing that

$$\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid c, w, x) U(c, y^*; w) = \mathbb{E}_Q \left[U(c, Y^*; w) \mid C = c, W = w, X = x \right], \\ \sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid c, w, x) U(c', y; w) = \mathbb{E}_Q \left[U(c', Y^*; w) \mid C = c, W = w, X = x \right].$$

The statement in the main text is the special case with $\mathcal{C} = \{0, 1\}$ and $\mathcal{C}^y = \{1\}$. □

Proof of Theorem 2.2

Lemma C.2. *The human DM's choices are consistent with expected utility maximization behavior at some strict preference utility function if and only if there exists strict preference utility functions U satisfying*

i. $P(Y = 1 \mid C = 1, W = w, X = x) \leq \frac{U(0,0;w)}{U(0,0;w) + U_w(1,1)}$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$ *with* $P(C = 1 \mid W = w, X = x) > 0$,

ii. $\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \overline{P}(Y = 1 \mid C = 0, W = w, X = x)$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$ *with* $P(C = 0 \mid W = w, X = x) > 0$.

Proof. This is an immediate consequence of applying Lemma C.1 to binary screening decisions with strict preferences.

For all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $P(C = 1 \mid W = w, X = x) > 0$, condition (i) in Lemma C.1 requires

$$P(Y^* = 1 \mid C = 1, W = w, X = x) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)}.$$

For all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $P(C = 0 \mid W = w, X = x) > 0$, condition (ii) in Lemma C.1 requires

$$\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq P(Y^* = 1 \mid C = 0, W = w, X = x).$$

Applying the bounds $\underline{P}(Y^* = 1 \mid C = 0, W = w, X = x) \leq P(Y^* = 1 \mid C = 0, W = w, X = x) \leq \overline{P}(Y^* = 1 \mid C = 0, W = w, X = x)$ then delivers the result. □

By Lemma C.2, the human DM's choices are consistent with expected utility maximization behavior if and only if there exists utility functions U satisfying

$$\max_{x \in \mathcal{X}^1(w)} P(Y = 1 \mid C = 1, W = w, X = x) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)}$$
$$\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}(Y = 1 \mid C = 0, W = w, X = x)$$

for all $w \in \mathcal{W}$. These inequalities are only non-empty if the stated conditions in Theorem 2.2. The characterization of the identified set of utility functions also follows from Lemma C.2. □

Proof of Proposition 3.1

The stated bounds follow from rearranging the bounds in Proposition B.1 in a binary screening decision, which are sharp as discussed in Appendix B.2. Sharp bounds on the marginal distribution of Y^* are given by, for each $z \in \mathcal{Z}$, $P(C = 1, Y = 1 \mid W = w, X = x, Z = z) \leq P(Y^* = 1 \mid W = w, X = x) \leq P(C = 1, Y = 1 \mid W = w, X = x, Z = z) + P(C = 0 \mid W = w, X = x, Z = z)$. This additionally delivers sharp bonds on the marginal distribution of Y^* conditional on any $z \in \mathcal{Z}$ since the instrument is assumed to be independent of the outcome of interest. The sharpness of the bounds on $P(C = 0, Y^* = 1 \mid W = w, X = x, Z = z)$ immediately follows since $P(C = 1, Y^* = 1 \mid W = w, X = x, Z = z) = P(C = 1, Y = 1 \mid W = w, X = x)$ is observed. □

Proof of Theorem 4.1

To prove this result, I first show that a series of modified revealed preference inequalities are necessary and sufficient to characterize the conditions under which the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs. I prove this result for the case in which the decision maker faces many possible choices $\mathcal{C}$ with $|\mathcal{C}| := N_c$ as in the proof of Theorem 2.1. The researcher only observes the outcome of interest if $C \in \mathcal{C}^y \subseteq \mathcal{C}$ for some known subset $\mathcal{C}^y$ and the researcher places bounds of the form: for each $c \in \mathcal{C} \setminus \mathcal{C}^y$ and $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists a known subset $\mathcal{B}_{c,w,x} \subseteq \Delta(\mathcal{Y})$ such that $P_{Y^*}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$, where $P_{Y^*}(\cdot \mid c, w, x) \in \Delta(\mathcal{Y})$ is the conditional distribution of the outcome given $C = c$,

$W = w, X = x$ as in the main text. In this more general setting, the definition of expected utility maximization behavior at inaccurate beliefs (Definition 6) extends directly.

Lemma C.3. *Assume $P_{Y^*}(\cdot \mid w,x) > 0$ for all $P_{Y^*}(\cdot \mid w,x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w,x))$. The decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs if and only if there exists a utility function $U \in \mathcal{U}$, prior beliefs $Q_{Y^*}(\cdot \mid w,x) \in \Delta(\mathcal{Y})$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ and $\widetilde{P}_{Y^*}(\cdot \mid c,w,x) \in \mathcal{B}_{c,w,x}$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$, $c \in \mathcal{C} \setminus \mathcal{C}^y$ satisfying for all $c \in \mathcal{C}$ and $c' \neq c$*

$$\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid w,x)\widetilde{P}_C(c \mid y^*,w,x)U(c,y^*;w) \geq \sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid w,x)\widetilde{P}_C(c \mid y^*,w,x)U(c',y;w),$$

where

$$\widetilde{P}_C(c \mid y^*,w,x) = \begin{cases} \frac{P_{C,Y^*}(c,y^* \mid w,x)}{\widetilde{P}_{Y^*}(y \mid w,x)} \text{ if } c \in \mathcal{C}^y \\ \frac{\widetilde{P}_{Y^*}(y^* \mid c,w,x)P_C(c \mid w,x)}{\widetilde{P}_{Y^*}(y \mid w,x)} \text{ if } c \in \mathcal{C} \setminus \mathcal{C}^y \end{cases}$$

and $\widetilde{P}_{Y^*}(y \mid w,x) = \sum_{c \in \mathcal{C} \setminus \mathcal{C}^y} \widetilde{P}_{Y^*}(y^* \mid c,w,x)P_C(c \mid w,x) + \sum_{c \in \mathcal{C}^y} P_{C,Y^*}(c,y^* \mid w,x)$.

Proof of Lemma C.3: Necessity To show necessity, we apply the same steps as the proof of necessity for Lemma C.1. First, by an analogous argument as given in the proof of necessity for Lemma C.1, it is without loss of generality to assume the private information $V \in \mathcal{V}$ has finite support. Second, following the same steps as the proof of necessity for Lemma C.1, I arrive at

$$\sum_{y \in \mathcal{Y}} Q_{C,Y^*}(c,y^* \mid w,x)U(c,y^*;w) \geq \sum_{y^* \in \mathcal{Y}} Q_{C,Y^*}(c,y^*,w,x)U(c',y^*;w).$$

Then, we immediately observe that $Q_{C,Y^*}(c,y^* \mid w,x) = Q_C(c \mid y^*,w,x)Q_{Y^*}(y^* \mid w,x) = \widetilde{P}_C(c \mid y^*,w,x)Q_{Y^*}(y^* \mid w,x)$, where the last equality follows via data consistency with inaccurate beliefs.

Proof of Lemma C.3: Sufficiency To show sufficiency, suppose that the conditions in Lemma C.3 are satisfied at some $\widetilde{P}_{Y^*}(\cdot \mid c,w,x) \in \mathcal{B}_{c,w,x}$ for all $c \in \mathcal{C} \setminus \mathcal{C}^y$, $(w,x) \in \mathcal{W} \times \mathcal{X}$ and some $Q_{Y^*}(\cdot \mid w,x) \in \Delta(\mathcal{Y})$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$.

Define the joint distribution $(W,X,C,Y^*) \sim \widetilde{P}$ according to $\widetilde{P}(w,x,c,y^*) = \widetilde{P}_C(c \mid y^*,w,x)Q_{Y^*}(y^* \mid w,x)P(w,x)$, where $\widetilde{P}_C(\cdot \mid y^*,w,x)$ is defined in the statement of the Lemma. Given the inequalities in the Lemma, we can construct a joint distribution $(W,X,V,C,Y^*) \sim Q$ to satisfy information set, expected utility maximization behavior and data consistency (Definition 3) for the constructed joint distribution $(W,X,C,Y^*) \sim \widetilde{P}$ following the same sufficiency argument as given in Lemma C.1. This constructed joint distribution $(W,X,V,C,Y^*) \sim Q$ will be an expected utility maximization representation under inaccurate beliefs.

As notation, define $\widetilde{P}_C(c \mid w,x)$ to be the probability of $C = c$ given $W = w, X = x$ and $\widetilde{P}_{Y^*}(y^* \mid c,w,x)$ to be the choice-dependent outcome probability given $C = c, W = w, X = x$ under the constructed joint distribution $(W,X,C,Y^*) \sim \widetilde{P}$. Let $v \in \mathcal{V} := \{1,\ldots,2^{N_c}\}$ uniquely index all possible subsets in the power set $2^{\mathcal{C}}$.

For each $(w,x) \in \mathcal{W} \times \mathcal{X}$, define $\mathcal{C}_{w,x} := \left\{c\colon \widetilde{P}(c \mid w,x) > 0\right\} \subseteq \mathcal{C}$ to be the set of choices selected with positive probability, and partition $\mathcal{C}_{w,x}$ into subsets that have identical constructed choice-dependent outcome probabilities. There are $\bar{V}_{w,x} \leq |\mathcal{C}_{w,x}|$ such subsets. Associate each subset in this partition with its associated index $v \in \mathcal{V}$. Denote the choice-dependent outcome probability associated with the subset labelled v by $\tilde{P}_{Y^*}(\cdot \mid v,w,x) \in \Delta(\mathcal{Y})$.

Define the random variable $V \in \mathcal{V}$ according to

$$Q(V = v \mid w,x) = \sum_{c\colon \widetilde{P}_{Y^*}(\cdot|c,w,x) = \widetilde{P}_{Y^*}(\cdot|v,w,x)} \widetilde{P}_C(c \mid w,x) \text{ if } v \in \mathcal{V}_{w,x},$$

$$Q(V = v \mid y^*,w,x) = \begin{cases} \frac{\widetilde{P}_{Y^*}(y^*|v,w,x)Q(V=v|w,x)}{Q_{Y^*}(y^*|w,x)} \text{ if } v \in \mathcal{V}_{w,x} \text{ and } Q_{Y^*}(y^* \mid w,x) > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Next, define the random variable $C \in \mathcal{C}$ according to

$$Q(C = c \mid v,w,x) = \begin{cases} \frac{\widetilde{P}_C(c|w,x)}{\sum_{\tilde{c}\colon \widetilde{P}_{Y^*}(\cdot|\tilde{c},w,x) = \widetilde{P}_{Y^*}(\cdot|v,w,x)} \widetilde{P}_C(\tilde{c}|w,x)} \text{ if } v \in \mathcal{V}_{w,x} \text{ and } \widetilde{P}_{Y^*}(\cdot \mid c,w,x) = \widetilde{P}_{Y^*}(\cdot \mid v,w,x) \\ 0 \text{ otherwise.} \end{cases}$$

Together, this defines the random vector $(W,X,Y^*,V,C) \sim Q$, whose joint distribution is defined as

$$Q(w,x,y^*,v,c) = P(w,x)Q_{Y^*}(y^* \mid w,x)Q_V(v \mid y^*,w,x)Q_C(c \mid v,w,x).$$

We now check that this construction satisfies information set, expected utility maximization and data consistency. First, information set is satisfied since $Q_{C,Y^*}(c,y^* \mid w,x,v) = Q_{Y^*}(y^* \mid w,x,v)Q_C(c \mid w,x,v)$ by construction. Next, for any $(w,x) \in \mathcal{W} \times \mathcal{X}$ and $c \in \mathcal{C}_{w,x}$, define $v_{c,w,x} \in \mathcal{V}_{w,x}$ to be the label satisfying $\widetilde{P}_{Y^*}(\cdot \mid c,w,x) = \widetilde{P}_{Y^*}(\cdot \mid c,w,x)$. For $\widetilde{P}_{C,Y^*}(c,y^* \mid w,x) > 0$, observe that

$$\begin{aligned}
&\widetilde{P}_{C,Y^*}(c,y^* \mid w,x) = \\
&\widetilde{P}_{Y^*}(y^* \mid c,w,x)\widetilde{P}_C(c \mid w,x) = \\
&Q_{Y^*}(y^* \mid w,x)\frac{Q_{Y^*}(y^* \mid v_{c,w,x},w,x)\sum_{\left\{\tilde{c}\colon \substack{\widetilde{P}_{Y^*}(\cdot|\tilde{c},w,x)= \\ \widetilde{P}_{Y^*}(\cdot|v,w,x)}\right\}} \widetilde{P}_C(\tilde{c} \mid w,x)}{Q_{Y^*}(y^* \mid w,x)} \frac{\widetilde{P}_C(c \mid w,x)}{\sum_{\left\{\tilde{c}\colon \substack{\widetilde{P}_{Y^*}(\cdot|\tilde{c},w,x)= \\ \widetilde{P}_{Y^*}(\cdot|v,w,x)}\right\}} \widetilde{P}_C(\tilde{c} \mid w,x)} = \\
&Q_{Y^*}(y^* \mid w,x)Q_V(v_{c,w,x} \mid y^*,w,x)Q_C(c \mid v_{c,w,x},w,x) = \\
&\sum_{v \in \mathcal{V}} Q_{Y^*}(y^* \mid w,x)Q_V(v \mid y^*,w,x)Q_C(c \mid v,w,x) = \sum_{v \in \mathcal{V}} Q(v,c,y^* \mid w,x).
\end{aligned}$$

Moreover, whenever $\widetilde{P}_{C,Y^*}(c,y^* \mid w,x) = 0$, $Q_{Y^*}(y^* \mid v_{c,w,x},w,x)Q_C(c \mid v_{c,w,x},w,x) = 0$. Therefore, data consistency holds (Definition 3) holds for the constructed joint distribution $(W,X,C,Y^*) \sim \widetilde{P}$. Since $\widetilde{P}_{C,Y^*}(c,y^* \mid w,x) = \widetilde{P}_C(c \mid y^*,w,x)Q_{Y^*}(y^* \mid w,x)$ by construction, $(W,X,V,C,Y^*) \sim \widetilde{Q}$ satisfies data consistency at inaccurate beliefs (Definition 6). Finally, for

$Q(C = c \mid V = v_{c,w,x}, W = w, X = x) > 0$,

$$Q(Y^* = y^* \mid V = v_{c,w,x}, W = w, X = x) =$$
$$\frac{Q(V = v_{c,w,x} \mid Y^* = y^*, W = w, X = x) Q(Y^* = y^* \mid W = w, X = x)}{Q(V = v_{c,w,x} \mid W = w, X = x)} =$$
$$\widetilde{P}(Y^* = y^* \mid C = c, W = w, X = x)$$

and $\widetilde{P}(C = c \mid W = w, X = x) > 0$. Therefore, using data consistency and the inequalities in Lemma C.3, we have that

$$\sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid v, w, x) U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} Q_{Y^*}(y^* \mid v, w, x) U(c, y^*; w),$$

where we used that $Q_{Y^*}(y^* \mid w, x)\tilde{P}_C(c \mid y^*, w, x) = Q_{C,Y^*}(c, y^* \mid w, x)$ by data consistency and $Q_{C,Y^*}(c, y^* \mid w, x)/\tilde{P}_C(c \mid W = w, X = x) = \tilde{P}_{Y^*}(y^* \mid c, w, x)$ by data consistency as well. Therefore, expected utility maximization is also satisfied.

Rewrite inequalities in Lemma C.3 in terms of weights:

Lemma C.4. *Assume $P_{Y^*}(\cdot \mid w, x) > 0$ for all $P_{Y^*}(\cdot \mid w, x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w, x))$. The decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs if and only if there exists a utility function $U \in \mathcal{U}$, $\widetilde{P}_{Y^*}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$, $c \in \mathcal{C} \setminus \mathcal{C}^y$, and non-negative weights $\omega(y^*; w, x)$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$, $y^* \in \mathcal{Y}$ satisfying*

i. For all $c \in \mathcal{C}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $P(c \mid w, x) > 0$, $c' \neq c$

$$\mathbb{E}_{\widetilde{P}}\left[\omega(Y^*; W, X) U(c, Y^*; W) \mid C = c, W = w, X = x\right] \geq$$
$$\mathbb{E}_{\widetilde{P}}\left[\omega(Y^*; W, X) U(c', Y^*; W) \mid C = c, W = w, X = x\right]$$

ii. For all $(w, x) \in \mathcal{W} \times \mathcal{X}$, $\mathbb{E}_{\widetilde{P}}[\omega(Y^*; W) \mid W = w, X = x] = 1$

where $\mathbb{E}_{\widetilde{P}}[\cdot]$ is the expectation under the joint distribution under $\widetilde{P}$ ad $\widetilde{P}$ is defined as

$$\widetilde{P}(w, x, c, y^*) = \begin{cases} P(w, x, c, y^*) \text{ if } c \in \mathcal{C}^y, \\ \widetilde{P}_{Y^*}(y^* \mid c, w, x) P(c, w, x) \text{ if } c \in \mathcal{C} \setminus \mathcal{C}^y. \end{cases}$$

Proof. Define $\widetilde{P}$ as in the statement of the lemma. Rewrite the condition in Lemma C.3 as: for all $c \in \mathcal{C}$ and $\tilde{c} \neq c$,

$$\sum_{y \in \mathcal{Y}} \frac{Q_{Y^*}(y^* \mid w, x)}{\widetilde{P}_{Y^*}(y^* \mid w, x)} \widetilde{P}_{C,Y^*}(c, y^* \mid w, x) U(c, y^*; w) \geq \sum_{y \in \mathcal{Y}} \frac{Q_{Y^*}(y^* \mid w, x)}{\widetilde{P}_{Y^*}(y^* \mid w, x)} \widetilde{P}_{C,Y^*}(c, y^* \mid w, x) U(c', y^*; w).$$

Notice that if $P_C(c \mid w, x) = 0$, then $\widetilde{P}_{C,Y^*}(c, y \mid w, x) = 0$. Therefore, the inequalities involving $c \in \mathcal{C}$ with $P_C(c \mid w, x) = 0$ are trivially satisfied. Next, inequalities involving $c \in \mathcal{C}$ with

$P_C(c \mid w, x) > 0$ can be equivalently rewritten as

$$\sum_{y^* \in \mathcal{Y}} \frac{Q_{Y^*}(y^* \mid w, x)}{\widetilde{P}_{Y^*}(y^* \mid w, x)} \widetilde{P}_{Y^*}(y^* \mid c, w, x) U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} \frac{Q_{Y^*}(y^* \mid w, x)}{\widetilde{P}_{Y^*}(y^* \mid w, x)} \widetilde{P}_{Y^*}(y^* \mid c, w, x) U(c', y^*; w).$$

The statement of Lemma C.4 follows by noticing that $\sum_{y^* \in \mathcal{Y}} \widetilde{P}_{Y^*}(y^* \mid c, w, x) \frac{Q_{Y^*}(y^*|w,x)}{\widetilde{P}_{Y^*}(y^*|w,x)} U(c, y^*; w) = \mathbb{E}_{\widetilde{P}}\left[\frac{Q_{Y^*}(y^*|w,x)}{\widetilde{P}_{Y^*}(y^*|w,x)} U(c, y^*; w)\right]$ and defining the weights as $\omega(y^*; w, x) = \frac{Q_{Y^*}(y^*|w,x)}{\widetilde{P}_{Y^*}(y^*|w,x)}$. □

Lemma C.4 implies Theorem 4.1 I now prove Theorem 4.1 as a consequence of Lemma C.4. Under the stated conditions, if the decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs and some strict preference utility function, Lemma C.4 implies that for all $(w, x) \in \mathcal{W} \times \mathcal{X}$

$$\omega(1; w, x) U(1, 1; w, x) P(Y^* = 1 \mid C = 1, W = w, X = x) \geq$$

$$\omega(0; w, x) U(0, 0; w, x) P(Y^* = 0 \mid C = 1, W = w, X = x),$$

$$\omega(0; w, x) U(0, 0; w, x) \widetilde{P}(Y^* = 0 \mid C = 0, W = w, X = x) \geq$$

$$\omega(1; w, x) U(1, 1; w, x) \widetilde{P}(Y^* = 1 \mid C = 1, W = w, X = x),$$

where $\omega(y^*; w, x) = \frac{\widetilde{Q}(y^*|w,x)}{\widetilde{P}(y^*|w,x)}$. Re-arranging these inequalities, we observe that

$$P(Y^* = 1 \mid C = 1, W = w, X = x) \leq \frac{\omega(0; w, x) U(0, 0; w, x)}{\omega(0; w, x) U(0, 0; w, x) + \omega(1; w, x) U(1, 1; w, x)}$$

$$\frac{\omega(0; w, x) U(0, 0; w, x)}{\omega(0; w, x) U(0, 0; w, x) + \omega(1; w, x) U(1, 1; w, x)} \leq \widetilde{P}(Y^* = 1 \mid C = 0, W = w, X = x).$$

The result then follows by applying the bounds on $\widetilde{P}(Y^* = 1 \mid C = 0, W = w, X = x)$ in a binary screening decision. □

Supplementary Materials to Identifying Prediction Mistakes in Observational Data

Ashesh Rambachan

August 26, 2021

D Extensions: Expected Utility Maximization Behavior

D.1 Treatment Assignment Problems

In this section, I extend the setting described in Section 2 of the main text to analyze treatment assignment problems, in which the decision maker selects between many choices and her choices have a causal effect on the outcome. This nest the analysis of screening decisions in the main text as a special case.

D.1.1 The Observable Data in Treatment Assignment Problems

The decision maker selects a choice $c \in \mathcal{C} := \{c_1, \ldots, c_{N_c}\}$ for each individual. Each individual is summarized by finite support characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$ as before, but now each individual is additionally associated with a vector of potential outcomes. The *potential outcome* $y_k := y(c_k) \in \mathcal{Y}$ is the outcome that would occur if the decision maker were to select choice c_k, where the support $\mathcal{Y}$ is finite. Let $\vec{y} := (y_1, \ldots, y_{N_c})$ be the vector of potential outcomes associated with each choice and $\vec{y}_{-k}$ be the vector of all potential outcomes except for the potential outcome associated with choice c_k. The random vector $(W, X, C, \vec{Y}) \sim P$ defined over the sample space $\mathcal{W} \times \mathcal{X} \times \mathcal{C} \times \mathcal{Y}^{N_c}$ summarizes the joint distribution of the characteristics, the decision maker's choices and potential outcomes across all individuals. I continue to assume $P(W = w, X = x) \geq \delta$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ for some $\delta > 0$ as in the main text.

The researcher observes the characteristics (W, X) for each individual as well as the decision maker's choice C. The missing data problem arises from the fundamental problem of causal inference (Holland, 1986). The researcher only observes the potential outcome associated with the choice selected by the decision maker, where $Y := \sum_{k=1}^{N_c} Y_k 1\{C = c_k\}$ is now the observable outcome. The observable choice-dependent outcome probabilities are $P(Y_k = y \mid C = c_k, W = w, X = x)$ for each $y \in \mathcal{Y}$, and $k = 1, \ldots, N_c$. The researcher does not observe the choice-dependent distribution of the full potential outcome vector $P(\vec{Y} = \vec{y} \mid C = c_k, W = w, X = x)$ for each $k = 1, \ldots, N_c$.

As notation, let $P_{\vec{Y}}(\vec{y} \mid c_k, w, x) := P(\vec{Y} = \vec{y} \mid C = c_k, W = w, X = x)$ and define $P_{\vec{Y}}(\cdot \mid c_k, w, x) \in \Delta(\mathcal{Y}^{N_c})$ denote the conditional distribution of $\vec{Y} \mid \{C = c_k, W = w, X = x\}$.

I assume that the researcher places bounds on the missing data. For each choice $c_k \in \mathcal{C}$ and observable characteristic $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists $\mathcal{B}_{c_k,w,x} \in \Delta(\mathcal{Y}^{N_c})$ satisfying $P_{\vec{Y}}(\cdot \mid c_k, w, x) \in \mathcal{B}_{c_k,w,x}$ and each $\tilde{P}_{\vec{Y}}(\cdot \mid c_k, w, x) \in \mathcal{B}_{c_k,w,x}$ satisfies $\sum_{\vec{y}_{-k}} \tilde{P}_{\vec{Y}}(\vec{Y}_{-k} = \vec{y}_{-k}, Y_k = y_k \mid C = c_k, W = w, X = x) = P(Y_k = y_k \mid C = c_k, W = w, X = x)$. Under such bounds, the choice-dependent distribution of the full potential outcome vector is partially identified.

Example: Medical Treatment A doctor decides what medical treatment to give a patient based on a prediction of how that medical treatment will affect the patient's health outcomes (Chandra and Staiger, 2007, 2011; Manski, 2017; Currie and Macleod, 2017; Abaluck et al., 2020; Currie and Macleod, 2020). For example, a doctor decides whether to give reperfusion therapy to an admitted patient that has suffered from a heart attack (Chandra and Staiger, 2020). The outcome $Y \in \{0, 1\}$ denotes whether the patient died within 30 days of admission and the choice $C \in \{0, 1\}$ denotes whether the patient was given reperfusion therapy within 12 hours of admission to the hospital. The potential outcomes Y_0, Y_1 are whether the patient would have died within 30 days of admission had the doctor not given or given reperfusion therapy to the patient respectively. The characteristics (W, X) summarize rich information about the patient that is available at the time of the patient's admission with a heart attack such as demographic information, collected vital signs and the patient's prior medical history. The researcher observes the doctor's propensity to give reperfusion therapy at each characteristic, $P(C = 1 \mid W = w, X = x)$. The researcher observes the 30-day mortality rate among patients that did not receive reperfusion therapy, $P(Y_0 = 1 \mid C = 0, W = w, X = x)$, and the 30-day mortality rate among patients that received reperfusion therapy, $P(Y_1 = 1 \mid C = 1, W = w, X = x)$. For patients that received reperfusion therapy, the researcher cannot observe their counterfactual mortality rate had the doctor not given reperfusion therapy, $P(Y_0 = 1 \mid C = 1, W = w, X = x)$. For patients that did not receive reperfusion therapy, the researcher cannot observe their counterfactual mortality rate had the doctor given them reperfusion therapy, $P(Y_1 = 1 \mid C = 0, W = w, X = x)$. ▲

Example: Credit Pricing Among approved loans, a loan officer decides what interest rate to charge based on a prediction of how different interest rates would affect the loan applicant's probability of default. The outcome $Y \in \{0, 1\}$ denotes whether the loan applicant would default on the loan and the choice $C \in \{c_1, \ldots, c_{N_c}\}$ denotes various interest rates that could be charged. The potential outcomes $(Y_1, \ldots, Y_{N_c})$ are whether the loan applicant would default under each possible interest rate. The characteristics (W, X) summarize the loan applicant's demographics and financial information such as income, savings and prior credit history. The researcher observes the loan officer's probability of charging a particular interest rate at each observable characteristic, $P(C = c_k \mid W = w, X = x)$, and the default rate associated with each charged interest rate $P(Y_k = 1 \mid C = c_k, W = w, X = x)$. The researcher cannot observe the counterfactual default rate had the loan officer charged a different interest rate, for example, $P(Y_{\tilde{k}} = 1 \mid C = c_k, W = w, X = x)$. ▲

D.1.2 Expected Utility Maximization Behavior and the Identification Result

The expected utility maximization model presented in Section 2 requires only minor changes to account for the introduction of potential outcomes. In particular, the utility function now specifies the payoff associated with each possible combination of choice and potential outcome vector at all possible observable characteristics.

Definition 7 (Utility Function in Treatment Assignment Problems)**.** The *utility function* $U \colon \mathcal{C} \times \mathcal{Y}^{N_c} \times \mathcal{W}$ specifies the payoff associated with each choice and vector of potential outcomes pair at each possible value of the observable characteristics. $U(c, \vec{y}; w)$ is the payoff associated with choice c and vector of potential outcomes $\vec{y}$ at characteristics $w \in \mathcal{W}$. Let $\mathcal{U}$ denote the set of feasible utility functions.

Private information is defined in the same way as Definition 2. I next extend Definition 3 in Section 2 to suitably account for the introduction of potential outcomes.

Definition 8 (Expected Utility Maximization Behavior in Treatment Assignment Problems)**.** The decision maker's choices are *consistent with expected utility maximization behavior* if there exists some utility function $U \in \mathcal{U}$ and joint distribution $(W, X, V, C, \vec{Y}) \sim Q$ satisfying

i. Information Set: $C \perp\!\!\!\perp \vec{Y} \mid W, X, V$ under Q.

ii. Expected Utility Maximization: For all $c \in \mathcal{C}$, $(w, x, v) \in \mathcal{W} \times \mathcal{X} \times \mathcal{V}$ such that $Q_C(c \mid w, x, v) > 0$,

$$\mathbb{E}_Q\left[U(c, \vec{Y}; W) \mid W = w, X = x, V = v\right] \geq \mathbb{E}_Q\left[U(c', \vec{Y}; W) \mid W = w, X = x, V = v\right]$$

for all $c' \neq c$, where $\mathbb{E}_Q[\cdot]$ denotes the expectation under Q.

iii. Data Consistency: For all $c_k \in \mathcal{C}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists $\widetilde{P}_{\vec{Y}}(\cdot \mid c_k, w, x) \in \mathcal{B}_{c_k,w,x}$ satisfying, for all $\vec{y} \in \mathcal{Y}^{N_c}$,

$$\widetilde{P}_{\vec{Y}}(\vec{y} \mid c_k, w, x) P(C_k = c, W = w, X = x) = Q(W = w, X = x, C = c_k, \vec{Y} = \vec{y}).$$

The decision maker's choices in a treatment assignment problem are consistent with expected utility maximization behavior if and only if there exists utility function that satisfies a series of revealed preference inequalities. The proof is the same as the proof of Theorem 2.1 as all that needs to be changed is that the outcome must now be defined as the full potential outcome vector.

Theorem D.1. *The decision maker's choices in a treatment assignment problem are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid c_k, w, x) \in \mathcal{B}_{c_k,w,x}$ for all $c_k \in \mathcal{C}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$ such that*

$$\mathbb{E}_Q\left[U(c_k, \vec{Y}; W) \mid C = c_k, W = w, X = x\right] \geq \mathbb{E}_Q\left[U(c', \vec{Y}; W) \mid C = c_k, W = w, X = x\right], \tag{8}$$

for all $c_k \in \mathcal{C}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $P(c_k \mid w, x) > 0$ and $c' \neq c_k$, where $\mathbb{E}_Q[\cdot]$ is the expectation under Q and

$$Q(W = w, X = x, C = c_k, \vec{Y} = \vec{y}) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid c_k, w, x) P(C = c_k, W = w, X = x).$$

D.1.3 Binary Treatment Assignment Problems

In Section 2.4 of the main text, I analyzed the testable implications of expected utility maximization behavior at accurate beliefs in binary screening decisions. I now show that these identification results also extend to treatment decisions with a binary outcome $\mathcal{Y} = \{0, 1\}$ by analyzing the implications of Theorem D.1 in a binary treatment assignment problem. I analyze the conditions under which the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs and a utility function that takes the form $U(c, \vec{y}; w) = b(y; w) - \lambda(c; w)$. This is an extended Roy model in which $b(y; w)$ is the benefit function that only depends on the realized

outcome and $\lambda(c;w)$ is the cost function that only depends on the choice. In this sense, this analysis relates to Henry, Meango and Mourifie (2020), which also studies a binary outcome extended Roy model under the assumption that utility function satisfies $U(0,\vec{y};w) = y_0, U(1,\vec{y};w) = Y_1 - \lambda(Y_1)$ for some function $\lambda(\cdot)$. Given this form of the utility function, it is without further loss of generality to normalize $b(0;w) = 0, \lambda(0;w) = 0$ for all $w \in \mathcal{W}$. I further focus on the case of strict preference inequalities in which $b(1;w) > 0$, $\lambda(1;w) > 0$ for all $w \in \mathcal{W}$.

Define $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon \pi_0(w,x) > 0\}$ and $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon \pi_1(w,x) > 0\}$. Let $P_{\vec{Y}}(y_0, y_1 \mid c, w, x) := P(Y_0 = y_0, Y_1 = y_1 \mid C = c, W = w, X = x)$ as shorthand.

Theorem D.2. *Consider a treatment decision with a binary outcome. The decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function $U(c,\vec{y};w) = b(y;w) - \lambda(c;w)$ if and only if for all $w \in \mathcal{W}$*

$$\max_{x \in \mathcal{X}^0(w)} \left\{ \underline{P}_{\vec{Y}}(0,1 \mid 0,w,x) - \underline{P}_{\vec{Y}}(1,0 \mid 0,w,x) \right\} \leq \min_{x \in \mathcal{X}^1} \left\{ \overline{P}_{\vec{Y}}(0,1 \mid 1,w,x) - \overline{P}_{\vec{Y}}(1,0 \mid 1,w,x) \right\},$$

where

$$\underline{P}_{\vec{Y}}(0,1 \mid 0,w,x) - \underline{P}_{\vec{Y}}(1,0 \mid 0,w,x) =$$
$$\min \left\{ \widetilde{P}_{\vec{Y}}(0,1 \mid 0,w,x) - \widetilde{P}_{\vec{Y}}(1,0 \mid 0,w,x) \colon \widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x} \right\}$$

and

$$\overline{P}_{\vec{Y}}(0,1 \mid 1,w,x) - \overline{P}_{\vec{Y}}(1,0 \mid 1,w,x) =$$
$$\max \left\{ \widetilde{P}_{\vec{Y}}(0,1 \mid 1,w,x) - \widetilde{P}_{\vec{Y}}(1,0 \mid 1,w,x) \colon \widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x} \right\}.$$

Proof. This result follows immediately from applying the inequalities in Theorem D.1 to the binary outcome case. Over $(w,x) \in \mathcal{W} \times \mathcal{X}$ such that $\pi_1(w,x) > 0$, the following inequality must be satisfied:

$$\frac{\lambda(1;w)}{b(1;w)} \leq P_{Y_1}(1 \mid 1,w,x) - P_{Y_0}(1 \mid 1,w,x).$$

Analogously, over $(w,x) \in \mathcal{W} \times \mathcal{X}$ such that $\pi_0(w,x) > 0$, the following inequality must be satisfied

$$\frac{\lambda(1;w)}{b(1;w)} \geq P_{Y_1}(1 \mid 0,w,x) - P_{Y_0}(1 \mid 0,w,x).$$

Moreover, notice that $P_{Y_1}(1 \mid 1,w,x) - P_{Y_0}(1 \mid 1,w,x) = P_{\vec{Y}}(0,1 \mid 1,w,x) - P_{\vec{Y}}(1,0 \mid 1,w,x)$ and $P_{Y_1}(1 \mid 0,w,x) - P_{Y_0}(1 \mid 0,w,x) = P_{\vec{Y}}(0,1 \mid 0,w,x) - P_{\vec{Y}}(1,0 \mid 0,w,x)$. The result is then immediate. □

This result immediately implies several negative results about the testability of expected utility maximization behavior that are analogous to those stated in the main text for a screening decision with binary outcomes. I state these as a corollary.

Corollary D.1. *Consider a treatment decision with a binary outcome. The decision maker's choices are consistent with expected utility maximization behavior some strict preference utility function $U(c,\vec{y};w) = b(y;w) - \lambda(c;w)$ if either:*

(i) All characteristics affect utility (i.e., $\mathcal{X} = \emptyset$) and $\underline{P}_{\vec{Y}}(0,1 \mid 0,w) - \underline{P}_{\vec{Y}}(1,0 \mid 0,w) \leq \overline{P}_{\vec{Y}}(0,1 \mid 1,w) - \overline{P}_{\vec{Y}}(1,0 \mid 1,w)$.

(ii) The researcher's bounds on the choice-dependent potential outcome probabilities are uninformative, meaning that for both $c \in \{0,1\}$, $\mathcal{B}_{c,w,x}$ is the set of all $\widetilde{P}_{\vec{Y}}(\cdot \mid c,w,x)$ satisfying $\sum_{y_{\tilde{c}} \in \mathcal{Y}} \widetilde{P}_{\vec{Y}}(\cdot, y_{\tilde{c}} \mid c,w,x) = P_{Y_c}(\cdot \mid c,w,x)$ for all $\widetilde{P}_{\vec{Y}}(\cdot \mid c,w,x) \in \mathcal{B}_{c,w,x}$.

Case (i) follows immediately from Theorem D.2. Case (ii) follows since under uninformative bounds on the missing data, $\overline{P}_{\vec{Y}}(0,1 \mid 1,w,x) - \overline{P}_{\vec{Y}}(1,0 \mid 1,w,x) = P(Y_1 = 1 \mid C = 1)$ and $\underline{P}_{\vec{Y}}(0,1 \mid 0,w,x) - \underline{P}_{\vec{Y}}(1,0 \mid 0,w,x) = -P(Y_0 = 1 \mid C = 0)$.

D.2 Continuous Characteristics

In this section, I extend the setting described in Section 2 to allow for the characteristics $X \in \mathcal{X}$ to be continuously distributed.

D.2.1 The Observable Data with Continuous Characteristics

I continue to assume that the outcome $Y^* \in \mathcal{Y}$, the choices $C \in \mathcal{C} := \{0,1\}$ and the characteristics $W \in \mathcal{W}$ are finite and now allow the characteristics $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ to be continuously distributed. The random vector $(W,X,C,Y^*) \sim P$ is defined over $\mathcal{W} \times \mathcal{X} \times \mathcal{C} \times \mathcal{Y}$ and summarizes the joint distribution of the characteristics, choices and outcome of interest. I assume the joint distribution P admits a density $p(w,x,c,y^*)$ that is continuous in x at each value $(w,c,y^*) \in \mathcal{W} \times \mathcal{C} \times \mathcal{Y}$ and satisfies $p(w,x) > 0$ for all $\mathcal{W} \times \mathcal{X}$.

The researcher observes the characteristics (W,X) and the human DM's choice C in each decision, but only observes the outcome Y^* if the human DM selected $C = 1$. Therefore, the researcher observes the joint distribution $(W,X,C,Y) \sim P$, where $Y := Y^* \cdot 1\{C = 1\}$.

The researcher places bounds on the unobservable choice-dependent outcome probabilities by specifying a family of conditional densities over (x,y^*) conditional on $W = w, C = c$, denoted by $\mathcal{B}_{c,w}$. Whenever the decision maker selects $C = 1$, the researcher observes (W,X,C,Y^*), and so $\mathcal{B}_{c=1,w}$ is a singleton that only contains the observable density $p(x,y^* \mid C = 1, W = w)$. Over the choice $c = 0$, the set $\mathcal{B}_{c=0,w}$ is a set of joint densities $\tilde{p}(x,y^* \mid W = w, C = 0)$ that satisfy $\sum_{\mathcal{Y}} \tilde{p}(X = x, Y^* = y^* \mid W = w, C = 0) = p(X = x \mid W = w, C = 0)$, where $p(X = x \mid W = w, C = 0)$ is the density $X \mid W = w, C = 0$.

D.2.2 Expected Utility Maximization Behavior and the Identification Result

The expected utility maximization model again requires minimal modification in order to account for the continuous characteristics. The definition of a utility function and private information is unchanged. I extend Definition 3 to ask whether there exists a random vector $(W,X,C,V,Y^*) \sim Q$ that admits a density $q(w,x,v,c,y^*)$ that is consistent with the observable data by simply replacing probability mass functions with the appropriate probability density function. Analogously, the characterization of expected utility maximization behavior also extends directly.

Theorem D.3. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ and $\tilde{p}(\cdot \mid C = 0, W = w) \in \mathcal{B}_{c=0,w}$ for all $w \in \mathcal{W}$ such that*

$$\mathbb{E}_Q\left[U(c,Y^*;W) \mid C = c, W = w, X = x\right] \geq \mathbb{E}_Q\left[U(c',Y^*;W) \mid C = c, W = w, X = x\right],$$

for all $c \in \{0,1\}$, $(w,x) \in \mathcal{W} \times \mathcal{X}$ *with* $P(C = c \mid W = w, X = x) > 0$ *and* $c' \neq c$, *where* $\mathbb{E}_Q[\cdot]$ *is the expectation under* Q *with density* q *satisfying*

$$q(W = w, X = x, C = 1, Y^* = y^*) = p(W = w, X = x, C = 1, Y^* = y^*),$$

$$q(W = w, X = x, C = 0, Y^* = y^*) = \widetilde{p}(X = x, Y^* = y^* \mid C = 0, W = w)p(C = 0, W = w).$$

Proof. The proof of this result is analogous to the proof of Theorem 2.1 Towards this, I first extend Lemma C.1 to the case with continuous characteristics (which analyzes the case with multiple choices and $|\mathcal{C}| = N_c$). Throughout, I write $p_{C,Y^*}(c, y^* \mid w, x) := P(C = c, Y^* = y^* \mid W = w, X = x)$ as shorthand, where notation such as $p_{X,Y^*}(x, y^* \mid w, c)$ is defined analogously.

Lemma D.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function* $U \in \mathcal{U}$ *that satisfies*

i. For all $c \in \mathcal{C}^y$, $(w,x) \in \mathcal{W} \times \mathcal{X}$ *and* $c' \neq c$,

$$\sum_{y^* \in \mathcal{Y}} p_{C,Y^*}(c, y^* \mid w, x)U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} p_{C,Y^*}(c, y^* \mid w, x)U(c', y^*; w).$$

ii. For all $c \in \mathcal{C} \setminus \mathcal{C}^y$ *and* $w \in \mathcal{W}$, *there exists* $\widetilde{p}_{C,Y^*}(\cdot \mid w, c) \in \mathcal{B}_{w,c}$ *such that*

$$\sum_{y^* \in \mathcal{Y}} \widetilde{p}_{C,Y^*}(c, y^* \mid w, x)U(c, y^*; w) \sum_{y^* \in \mathcal{Y}} \widetilde{p}_{C,Y^*}(c, y^* \mid w, x)U(c', y^*; w)$$

for all $x \in \mathcal{X}$ *and* $c' \neq c$, *where* $\widetilde{p}_{C,Y^*}(c, y^* \mid w, x) = \widetilde{p}_{X,Y^*}(x, y^* \mid w, c)p_{W,C}(w, c)/p_{W,X}(w, x)$.

Proof of Necessity for Lemma D.1: The proof of necessity follows the same argument as the proof of necessity of Lemma C.1 below by replacing the probability mass function Q with the density q. □

Proof of Sufficiency for Lemma D.1: The proof of sufficiency follows the proof of sufficiency of Lemma C.1 below by again simply replacing all probability mass functions with the appropriate density function. □

Theorem D.3 then follows directly from Lemma D.1 by considering the special case with $\mathcal{C} = \{0,1\}$ and $\mathcal{C}^y = \{1\}$. □

Theorem D.3 can be applied to binary screening decisions in which the characteristics $X \in \mathcal{X}$ are continuously distributed. In a binary screening decision, the bounds on the unobservable choice-dependent outcome probability $\mathcal{B}_{c=0,w}$ are simply joint densities $\tilde{p}(x, Y^* = 0 \mid W = w, C = 0)$, $\tilde{p}(x, Y^* = 1 \mid W = w, C = 0)$ that are continuous in $x \in \mathcal{X}$ and satisfy $p(x, Y^* = 0 \mid W = w, C = 0) + p(x, Y^* = 1 \mid W = w, C = 0) = p(x \mid W = w, C = 0)$. From Theorem D.3, the decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function U if and only if for all $w \in \mathcal{W}$ there exists $U(0,0;w) <$

$0, U(1,1;w) < 0$ satisfying

$$\sup_{x \in \mathcal{X}^1(w)} p(Y^* = 1 \mid C = 1, W = w, X = x) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \tag{9}$$

$$\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \inf_{x \in \mathcal{X}^0(w)} \overline{p}(Y^* = 1 \mid C = 0, W = w, X = x), \tag{10}$$

where $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon p(C = 0 \mid W = w, X = x) > 0\}$, $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon p(C = 1 \mid W = w, X = x) > 0\}$ and $\overline{p}(Y^* = 1 \mid C = 0, W = w, X = x)$ is the upper bound on $\tilde{p}(x, Y^* = 1 \mid W = w, C = 0)/p(x \mid W = w, C = 0)$ over densities satisfying the bounds $\mathcal{B}_{c=0,w}$. This provides a sharp characterization of the identified set of strict preference utility functions in terms of "intersection bounds." Therefore, researchers seeking to test whether a decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs can leverage inference procedures developed in, for example, Chernozhukov, Lee and Rosen (2013).

Finally, Theorem D.3 can also be simplified through dimension reduction over the continuously distributed characteristics $X \in \mathcal{X}$. Consider functions $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ for each $w \in \mathcal{W}$ that partition the characteristic space into level sets $\{x \in \mathcal{X} \colon D_w(x) = d\}$. In a binary screening decision, if the human DM's choices are consistent with expected utility maximization behavior at some strict preference utility function U that satisfies an exclusion restriction on the characteristics X, then

$$\max_{d \in \{1, \ldots, d_w\}} P(Y = 1 \mid C = 1, W = w, D_w(X) = d) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \tag{11}$$

$$\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \overline{P}(Y = 1 \mid C = 1, W = w, D_w(X) = d) \tag{12}$$

holds for all $w \in \mathcal{W}$. In this sense, the identification results for the case in which $X \in \mathcal{X}$ has finite support can be interpreted as a series of implied revealed preference inequalities that must be satisfied if the decision maker's choices are consistent with expected utility maximization behavior over the underlying continuously distributed characteristics.

E Extensions: Constructing Bounds on the Missing Data

E.1 Constructing Bounds on the Missing Data with Direct Imputation

The simplest empirical strategy for constructing bounds on the unobservable choice-dependent outcome probabilities is "direct imputation." In a binary screening decision, direct imputation uses the observable $P(Y = 1 \mid C = 1, W = w, X = x)$ to bound the unobservable $P(Y = 1 \mid C = 0, W = w, X = x)$.

Assumption E.1. For each $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $0 < P(C = 1 \mid W = w, X = x) < 1$, there exists $\kappa_{w,x} \geq 0$ satisfying

$$P(Y^* = 1 \mid C = 1, W = w, X = x) \leq P(Y^* = 1 \mid C = 0, W = w, X = x)$$
$$P(Y^* = 1 \mid C = 0, W = w, X = x) \leq (1 + \kappa_{w,x}) P(Y^* = 1 \mid C = 1, W = w, X = x).$$

The parameter $\kappa_{w,x} \geq 0$ specifies how different the unobservable choice-dependent outcome prob-

ability may be relative to the observable choice-dependent outcome probability. In pretrial release decisions, setting $\kappa_{w,x} = 1$ means that the researcher is willing to assume that the conditional probability of pretrial misconduct among detained defendants is no more than two times the conditional probability of pretrial misconduct among release defendants. Such bounding assumptions are used in, for example, Kleinberg et al. (2018), and Jung et al. (2020*a*).

In practice, the researcher may wish to test whether the decision maker is making mistakes under various choices of the parameter $\kappa_{w,x}$, and thereby conduct a sensitivity analysis of how robust the behavioral conclusions are to various assumptions about the unobservable choice-dependent outcome probabilities.

Finally, Assumption E.1 has a natural interpretation under the expected utility maximization model. The parameter $\kappa_{w,x}$ bounds the average informativeness of the decision maker's private information $V \in \mathcal{V}$.

Proposition E.1. *Consider a binary screening decision and suppose Assumption E.1 holds.*

If the decision maker's choices are consistent with expected utility maximization behavior at some private information $V \in \mathcal{V}$ and joint distribution $(W, X, V, C, Y^*) \sim Q$, then for each $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $0 < P(C = 1 \mid W = w, X = x) < 1$ and $0 < P(Y^* = 1 \mid C = 1, W = w, X = x) < 1$

a. $1 \leq \frac{Q(C=0|Y^*=1,W=w,X=x)/Q(C=1|Y^*=1,W=w,X=x)}{Q(C=0|W=w,X=x)/Q(C=1|W=w,X=x)} \leq 1 + \kappa_{w,x}$,

b. $1 - \kappa_{w,x} \frac{P(Y^*=1|C=1,W=w,X=x)}{P(Y^*=0|C=1,W=w,X=x)} \leq \frac{Q(C=0|Y^*=0,W=w,X=x)/Q(C=1|Y^*=0,W=w,X=x)}{Q(C=0|W=w,X=x)/Q(C=1|W=w,X=x)} \leq 1.$

Proof. Notice that

$$\frac{Q(C = 0 \mid Y^* = 1, W = w, X = x)/Q(C = 1 \mid Y^* = 1, W = w, X = x)}{Q(C = 0 \mid W = w, X = x)/Q(C = 1 \mid W = w, X = x)}$$

$$= \frac{Q(Y^* = 1 \mid C = 0, W = w, X = x)}{Q(Y^* = 1 \mid C = 1, W = w, X = x)}.$$

Since the decision maker's choices are consistent with expected utility maximization behavior, $(W, X, V, C, Y^*) \sim Q$ satisfies the data consistency condition in Definition 3 at some $\tilde{P}_{Y^*}(\cdot \mid c = 0, w, x)$ satisfying the bounds in Assumption E.1 for each $(w, x) \in \mathcal{W} \times \mathcal{X}$. Therefore, $Q(Y^* = 1 \mid C = 0, W = w, X = x) = \tilde{P}(Y^* = 1 \mid C = 0, W = w, X = x)$ and it immediately follows that

$$\frac{Q(Y^* = 1 \mid C = 0, W = w, X = x)}{Q(Y^* = 1 \mid C = 1, W = w, X = x)}$$

$$= \frac{\tilde{P}(Y^* = 1 \mid C = 0, W = w, X = x)}{P(Y^* = 1 \mid C = 1, W = w, X = x)} \in [1, 1 + \kappa_{w,x}]$$

under Assumption E.1. This proves (a). To show (b), notice that the bounds in Assumption E.1 imply that

$$P(Y^* = 0 \mid C = 1, W = w, X = x) - \kappa_{w,x} P(Y^* = 1 \mid C = 1, W = w, X = x)$$

$$\leq P(Y^* = 0 \mid C = 0, W = w, X = x) \leq P(Y^* = 0 \mid C = 1, W = w, X = x).$$

Moreover, as before, we also have that

$$\frac{Q(C=0 \mid Y^*=0, W=w, X=x)/Q(C=1 \mid Y^*=0, W=w, X=x)}{Q(C=0 \mid W=w, X=x)/Q(C=1 \mid W=w, X=x)}$$

$$= \frac{\tilde{P}(Y^*=0 \mid C=0, W=w, X=x)}{P(Y^*=0 \mid C=1, W=w, X=x)}$$

(b) then follows immediately.

□

The direct imputation bounds imply bounds on the relative odds ratio of the decision maker's choice probabilities conditional on the outcome and the characteristics relative to their choice probabilities conditional on only the characteristics. This places a bound on the average informativeness of the decision maker's private information under the expected utility maximization model, since

$$Q(C=1 \mid Y^*=1, W=w, X=x) = \mathbb{E}_Q\left[Q(C=1 \mid V=v, W=w, X=x) \mid Y^*=1, W=w, X=x\right]$$
$$Q(C=1 \mid Y^*=0, W=w, X=x) = \mathbb{E}_Q\left[Q(C=1 \mid V=v, W=w, X=x) \mid Y^*=0, W=w, X=x\right]$$

under the information set condition in Definition 3. In this sense, the direct imputation bounds are related to classic approaches for modelling violations of unconfoundedness in causal inference such as Rosenbaum (2002), which model violations of unconfoundedness by postulating that there exists some unobserved characteristics V that governs selection and places bounds on the magnitude of the relative odds ratio of the propensity score conditional on V and the observable characteristics versus the propensity score conditional on just the observable characteristics. See Imbens (2003), which develops a tractable parametric model for such a violation of unconfoundedness in a treatment assignment problem. Kallus, Mao and Zhou (2018) and Yadlowsky et al. (2020) derive bounds on the conditional average treatment effect and average treatment effect under related models for violations of unconfoundedness, and provide methods for inference on the derived bounds.

E.2 Constructing Bounds on the Missing Data with Proxy Outcomes

In some empirical applications, the researcher may observe an additional proxy outcome $\tilde{Y} \in \tilde{\mathcal{Y}}$, which does not suffer from the missing data problem but is correlated with the outcome $Y^* \in \mathcal{Y}$. By specifying bounds on the relationship between the proxy outcome $\tilde{Y}$ and the outcome Y^*, the researcher may construct bounds on the unobservable choice-dependent outcome probabilities.

Proxy outcomes are quite common in medical testing and loan approval settings. For example, Mullainathan and Obermeyer (2020) observe each patient's longer term health outcomes regardless of whether a stress test for a heart attack was conducted during a particular emergency room visit. A patient's longer term health outcomes are related to whether the patient actually had a heart attack, no matter the testing decisions of doctors. Similarly, Chan, Gentzkow and Yu (2020) observe whether each patient had a future pneumonia diagnosis within one week of an initial examination, regardless of whether a doctor ordered an MRI at the initial examination. Future pneumonia diagnoses may be a useful proxy for whether the doctor failed to correctly diagnose pneumonia during the initial examination. In mortgage approvals, Blattner and Nelson (2021) observe each loan ap-

plicant's default performance on other credit products, regardless of whether each loan applicant was approved for a mortgage. A loan applicant's default performance on other credit products is related with whether they would have defaulted on the mortgage.

Assumption E.2 (Proxy Outcomes)**.** There exists a proxy outcome $\tilde{Y} \in \tilde{\mathcal{Y}}$, and the researcher observes the joint distribution $(W, X, C, \tilde{Y}, Y^* \cdot 1\{C = 1\}) \sim P$. Assume $P(W = w, X = X, \tilde{Y} = \tilde{y}) > 0$ for all $(w, x, \tilde{y}) \in \mathcal{W} \times \mathcal{X} \times \tilde{\mathcal{Y}}$.

For simplicity, I focus on using proxy outcomes to construct bounds on the unobservable choice-dependent outcome probabilities in a binary screening decision. Over decisions in which the human DM selected $C = 1$, the researcher observes the joint distribution of the proxy outcome and the outcome $P(\tilde{Y} = \tilde{y}, Y^* = y^* \mid C = 1, W = w, X = x)$. By placing assumptions on how the joint distribution of the proxy outcome and the outcome conditional on $C = 0$ is bounded by the observable joint distribution of the proxy outcome and the outcome conditional on $C = 1$, the researcher can construct bounds on the unobservable choice-dependent outcome probabilities of the form given in Assumption 2.1.

Assumption E.3 (Proxy Bounds)**.** For each $(w, x, \tilde{y}) \in \mathcal{W} \times \mathcal{X} \times \tilde{\mathcal{Y}}$ satisfying $0 < P(C = 1 \mid \tilde{Y} = \tilde{y}, W = w, X = x) < 1$, there exists bounds $\underline{\kappa}_{\tilde{y},w,x}, \overline{\kappa}_{\tilde{y},w,x} \geq 0$ satisfying

$$P(Y^* = 1 \mid \tilde{Y} = \tilde{y}, C = 1, W = w, X = x) - \underline{\kappa}_{\tilde{y},w,x} \leq P(Y^* = 1 \mid \tilde{Y} = \tilde{y}, C = 0, W = w, X = x),$$
$$P(Y^* = 1 \mid \tilde{Y} = \tilde{y}, C = 0, W = w, X = x) \leq P(Y^* = 1 \mid \tilde{Y} = \tilde{y}, C = 1, W = w, X = x) + \overline{\kappa}_{\tilde{y},w,x}.$$

Proposition E.2. *Consider a binary screening decision in which Assumptions E.2-E.3 hold. For each* $(w, x) \in \mathcal{W} \times \mathcal{X}$*,*

$$\sum_{\tilde{y} \in \tilde{\mathcal{Y}}} \left\{ P(Y^* = 1 \mid \tilde{Y} = y, C = 1, W = w, X = x) - \underline{\kappa}_{\tilde{y},w,x} \right\} P(\tilde{Y} = y \mid C = 0, W = w, X = x) \leq$$
$$P(Y^* = y \mid C = 0, W = w, X = x) \leq$$
$$\sum_{\tilde{y} \in \tilde{\mathcal{Y}}} \left\{ P(Y^* = 1 \mid \tilde{Y} = y, C = 1, W = w, X = x) + \overline{\kappa}_{\tilde{y},w,x} \right\} P(\tilde{Y} = y \mid C = 0, W = w, X = x)$$

The advantage of this approach is that it may be easier for the researcher to express domain-specific knowledge through the use of proxy outcomes. For example, in the mortgage approvals setting in Blattner and Nelson (2021), the proxy bounds summarize the extent to which the mortgage default rate among accepted applicants that also defaulted on other credit products differs from the counterfactual mortgage default rate among rejected applicants that also defaulted on other credit products. In the medical testing setting in Mullainathan and Obermeyer (2020), the proxy bounds summarize the extent to which heart attack rate among tested patients that went on to die within 30 days of their emergency room visit differs from the heart attack rate among untested patients that went on to die within 30 days of their emergency room visit.