A Quick Note

Thank you very much for your interest in this paper. This is a very early and actively evolving draft which I'm using to record and summarize my general thoughts and observations regarding some interrelated topics, namely judge decisions, monotonicity, and inferences that can be made by comparing outcomes and decisions across subjects randomly assigned to different judges. My observations are somewhat disperse, and I believe that the points made in sections 2, 3, and 4 are conceptually separable. I am therefore particularly interested in feedback on the ideal way to package these observations (in the form of a single article or multiple articles) and, because this is my first time writing about these topics, I would particularly welcome feedback on pieces of the literature to which I should be connecting my findings. But, of course, I'm excited to receive all kinds of other feedback, too.

Thank you, and I'm looking forward to interacting with you!
– Murat

Endogenous Judge Decision Quality, Monotonicity, and Treatment Effects*

Murat C. Mungan†

November 28, 2022

Abstract

I consider a setting where judges sort unobservable types into two groups after incurring costs to increase the quality of their decisions. Judges who differ in their abilities, modeled as the marginal costs of quality investments, differ in their false positive and true positive categorization rates. I identify signal properties which lead judge decisions to violate monotonicity, as defined in instrumental variable designs exploiting variations in judge propensities in ways that are not detectable by tests that focus on variations across observable characteristics. I discuss inference problems when monotonicity is violated as well as misleading policy claims that can be based on findings even when monotonicity is not violated due to judges adjusting their quality investments in response to policy changes. These problems reveal the importance of distinguishing between average treatment effects by type, which I show can be identified when unobservables that are not solely a function of subjects' types have impacts on judge decisions and treatment effects by type which are orthogonal. I propose a method to test this condition.

Keywords: Judge decisions, monotonicity, instrumental variables, local average treatment effects, average treatment effects, judge design.

1 Introduction

One method that has been used frequently to sidestep endogeneity problems in empirical analyses is to use an instrument that it is correlated with the independent variable of interest, but not the dependent variable. In the law and economics literature, researchers have turned to judge propensities, e.g., judges' tendencies to return harsh verdicts, as an instrument to study the relationship between judge decisions and other outcomes of interest, e.g., recidivism. This

*I thank Adam Chilton, Bruce Kobayashi, Philip Marx, Naci Mocan, Jack Mountjoy, Adriana Robertson, Kyle Rozema, and participants at the University Of Chicago Law and Economics workshop for valuable comments and suggestions.

†Professor, Antonin Scalia Law School at George Mason University. e-mail: mmungan@gmu.edu.

approach has gained much popularity, and many economists refer to it simply as the 'judge design' approach. The types of inferences that can be made by using this approach, and the conditions under which it will return valid estimates of local average treatment effects (LATEs), are well studied. What is less studied, quite surprisingly, is the behavior of the component that is central to this type of analysis, namely the behavior of judges, which is the topic of this article.

Judges not only occupy a central role in these studies, but how judge decisions relate to each other affects the validity of the estimates obtained from them. One particular condition that is assumed to hold to guarantee valid estimates is a type of monotonicity across judge decisions. Although there are many variations of the general concept of monotonicity (e.g., strict, average, or probabilistic) the intuition behind them is similar: if a judge is stricter than another judge, then she must, at least in expectation, treat any given person stricter than the other judge. This assumption sounds intuitive, especially given the dominant narrative in this literature that what drives variation across judge propensities are their preferences: different judges have different values which cause them to select different decision criteria. However, as noted in the literature, judges may (sometimes due to non-legitimate reasons) have multi-dimensional preferences, which may cause monotonicity violations: although judge A may be stricter than judge B on average, she may be more lenient towards certain groups of people (e.g., as in the case of racial or gender discrimination).

This type of preference-based violation of monotonicity has been noted extensively in the literature. What has received less attention is whether judges may violate monotonicity because the *quality* of their decisions differ. To illustrate this type of variation, consider the simple hypothetical where a judge needs to decide whether to convict a defendant.[1] Suppose judge A randomizes and convicts half of the defendants she is assigned to, whereas judge B makes more accurate decisions which leads her to convict 80% of the truly guilty defendants assigned to her and convict only 10% of the truly innocent defendants assigned to her. This clearly constitutes a monotonicity violation.

This type of quality-based monotonicity violation is not driven by differences in judges' preferences: judges A and B in the above example may perceive similar costs from making erroneous decisions, and yet display very different decision profiles, simply because they differ in their ability to distinguish between different types of defendants. Perhaps more importantly, quality based monotonicity violations have testable implications which can be exploited to detect them: If there are no quality-based violations of monotonicity, then judges with high true positives would have to have high propensities, since their false positives would also need to be high. This implication can be tested, when the researcher can observe judges' miss-rates, e.g., granting bail to a person who later commits pre-trial misconduct, and measuring the correlation between them and judges' propensities.

Using the above-described test, Chan et al. (2022) have recently shown that

[1] I use this hypothetical to convey intuitions in a simple manner. In many cases in the United States, judges do not (at least exclusively) decide on convictions. However, this happens in other countries that do not use juries as well as in bench trials in the United States.

the decisions of the radiologists (to diagnose a patient with pneumonia based on X-Rays) whom they have studied clearly violate monotonicity, and due to differences in the quality of their decisions. In fact, Chan et al. demonstrate that there is a clear negative correlation between the true positive and false positive rates of the radiologists whom they have studied. This raises two important and interrelated preliminary questions. First, if there is a strong relationship between these two rates among one type of binary classifiers, namely radiologists, can one expect this relationship to exist among judges, as well? Second, what type of decision based dynamic may explain this type of relationship?

Here, I consider a model wherein judges make two interrelated decisions to return verdicts. First, they decide how much time and effort to devote to a particular case to increase the quality of their decisions. This investment enhances the average quality of their decisions by allowing them to review and interpret more relevant evidence, in addition to the basic facts of the case that is presented to them which serves as a signal of the defendant's type. Subsequently, based on all the information they have, they return a verdict to balance the costs of erroneous decisions. The second stage of this decision framework is the usual one, and has been considered by many others. The first stage of this analysis, to the best of my knowledge, is missing in the literature, and allows me to endogenize the quality of a judge's decision making.

In this context, judges with lower marginal costs of reviewing additional pertinent information can be interpreted as being more skilled. When judges differ in their skill levels, they may be associated with different decision-profiles. In fact, judges with greater skill naturally make higher quality decisions, because the cost of doing so is lower for them. However, what is *a priori* unclear is how higher decision quality is related to true and false positives. A judge who acquires superior information may decide to reduce one type of judicial error at the expense of another, or may decide to reduce both types of error by more modest amounts. Building on Lundberg and Mungan (2022), I show (in section 2.2.) that the judge will choose to reduce both types of error when the signal generating process possesses some intuitive properties. Quite importantly, this implies a violation of monotonicity: a judge with greater skill than another is associated with a higher true positive and a lower false positive rate. Thus, skill variation across judges can explain the types of monotonicity violations documented in Chan et al. (2022), which is a dynamic that is to be expected any time a binary classifier (whether it is a judge or a radiologist) must make a decision, and reviewing information is costly.

Although this type of monotonicity violation appears natural, its consequences with respect to the validity of inferences made through studies using judge propensities as instruments is quite severe. The literature notes that monotonicity violations can cause LATEs to be invalid, and even cause them to have the wrong sign (see, e.g., Angrist et al. 1996). This possibility is particularly relevant when monotonicity violations are caused due to quality differences in judge decisions and when treatment effects differ across subjects' types (e.g., truly guilty or innocent). A simple example can be used to demonstrate this point. Consider again judges A (50% conviction rate across the board) and B

(80% conviction rate for the truly guilty and 10% conviction rate for the truly innocent), and suppose that the proportion of truly guilty people is 50%. In this case, judge A will have a higher conviction rate (namely 50%) than judge B (namely 45%). Next, suppose that the conviction of truly innocent people has no impact on their tendencies to commit future offenses, but the conviction of the guilty *reduces* their propensities to commit future offenses. Given random judge assignment, this would cause the average future crime commission rate of defendants who have been assigned to judge A to have a higher crime commission rate than judge B. Thus, a study that uses judge propensities as an instrument would suggest that the average effect of a conviction on marginal defendants (i.e., those whose verdicts would be different under judges A and B) is positive. Yet, neither defendant type (i.e., guilty or innocent) experiences an increase in their tendencies to commit crimes in the future, and thus the true LATE is negative. I formalize this point in section 2.3., and identify conditions under which this type of sign reversal can occur. The analysis reveals that when the true LATE is negative, the estimated LATE may nevertheless be positive when true (resp. false) positives generate negative treatment effects when judges with higher propensities are associated with lower true (resp. false) positives. This highlights how categorizing and identifying treatment effects by defendant type (e.g., guilty, innocent) can be important.

Another problem may arise with endogenous judge decision quality, which relates to the interpretation, rather than the validity, of LATEs. For instance, a study might correctly find (i.e., in a case where there are no monotonicity violations) that the LATEs of convictions on the future criminal activity of defendants is positive. A seemingly plausible policy implication of this finding is that reducing judges' propensities to convict may reduce the future criminal activity of defendants. However, a policy that raises the costs of convictions to judges not only affects their preferences, but may also alter the amount of information they acquire to make decisions in any given case. I show that this will in fact cause judges to *reduce* the amount of information they acquire in cases where the acquisition of information causes judges to reduce their false positives. This highlights a particular manner in which policy recommendations based on valid LATEs may nevertheless generate unintended consequences due to the endogeneity of the information acquisition process of judges. The analysis of this problem (in section 2.4.) reveals that it emerges only when the true and false positives are associated with treatment effects of opposite signs.

This observation naturally leads to the question of whether there are any reasons, or any plausible mechanisms, which may cause average treatment effects to have opposite signs by type. In section 3, I propose a very simple extension of Miceli, Segerson, and Earnhart's (2022) model of specific deterrence, which provides a rationale for why this may happen in some contexts. I consider a setting where only some people who commit crimes along with some people who do not commit crimes are nevertheless punished (defined as the *positive* event), such that there are false and true positives. People who experience true positives (resp. false negatives) update their beliefs about the probability of receiving punishment upon committing crimes upwards (resp. downwards).

Similarly, people who experience false positives (resp. true negatives) update their beliefs about the probability of being punished despite not committing crime upwards (resp. downwards). Thus, people who commit crimes and are punished perceive higher expected opportunity costs associated with committing crimes compared to those who evade punishment despite committing crime, and thus commit crimes less frequently in the future. This corresponds to a negative treatment effect for truly guilty people when the outcome of interest is defendants' future crime commission rates. On the other hand, people who refrain from committing crime perceive a lower opportunity cost of committing future crimes when they are punished compared to when they are not, because their beliefs about the probability of being punished despite not committing crime is higher. This results in a positive treatment effect. Of course, this is not to suggest that treatments may not cause other partial positive effects (e.g., due to stigmatization (Mungan 2017a)) or other partial negative effects (e.g., due to repeat offenses being punished more severely Mungan (2017b)) across for both the truly guilty and innocent. This analysis is provided to illustrate how treatments may generate dynamics that go in opposite directions by type, for instance, due to asymmetric impacts of learning by subjects.

Given the inference and interpretation problems that may arise when treatment effects differ by types, and given theoretical reasons for why these effects may in fact differ, it becomes important to test for decision quality-based monotonicity violations and also to identify average treatment effects by type, separately. Therefore, in section 4, I first discuss the relationship between strong (*Global Monotonicity*) and weak (*Across Type Monotonicity*) variants of monotonicity. Across type monotonicity can be tested by using the approach proposed by Chan et al. (2022), but it is difficult to identify testable implications of instances where global monotonicity is violated despite across type monotonicity being satisfied. This is an important problem, because across type monotonicity is insufficient for valid LATEs. Thus, I specify a condition under which treatment effects associated with the two types of positives can be identified separately, even when neither monotonicity condition holds. The condition requires that the only unobservables that affect judges' decisions be generated solely by the true types of the defendants, and not other considerations. Although this condition may be satisfied in some settings, it may be too strong in others. Therefore, I identify another, orthogonality, condition, which requires that unobservables which are not solely produced by defendant types be independent from treatment effects. I show that these two conditions have very similar and testable implications, and propose a method to detect violations of them. I then explain how average treatment effects by type can be calculated when one of these two conditions hold –even when judge decisions are non-monotonic. The second part of this section includes hypothetical examples illustrating both the methods and tests developed in the first part.

In section 5, I provide concluding remarks regarding the crucial role of judges' decision making processes and implications regarding the importance of testing for quality-based violations of monotonicity.

2 Decision Quality, Monotonicity, and Inference Problems

In this section I model judge decisions as simple Bayesian decision problems wherein they choose a decision criterion to trade-off two possible types of decision errors. In this model, judges observe some *evidence* which is implerfectly informative of defendants' types, and choose a cut-off likelihood ratio (corresponding to a critical type of evidence) as their decision criterion (explained in subsection 1). In subsection 2, I consider judges who exogenously differ in their skills, which are captured through the likelihood with which they are able to observe additional information, called a *signal* -to distinguish it from the evidence described above- that is relevant to their decision making. I conduct comparative statics to identify how judge decisions, and in particular the errors committed by them, are affected by their ability. I identify broad sufficient conditions relating to evidence generation processes under which judge decisions are non-monotonic: True positives are increasing and false negatives are decreasing in judges' skills. In subsection 3, I explain how this type of non-monotonic behavior causes local average treatment effects (LATE) estimates obtained through intsrumental variables methods using judge designs to be invalid. In subsection 4, I show that even when judge decisions are monotonic, heterogeneity in judge skills can cause inferences regarding judge decision criteria, and therefore discriminatory practices, to be invalid. Finally, in subsection 5, I endogenize judge decision quality, by considering costly investments made by judges to increase the quality of their decisions. This analysis reveals that even when judge decisions are monotonic recommendations based on valid LATEs may nevertheless produce the opposite of the intended policy goals, because they may impact judges' investments in better decisions.

2.1 Judge's Decision Problem

A judge is tasked with returning *verdicts* $v \in \{p, n\}$ for subjects who are of one of two *types* $T \in \{P, N\}$ unobservable to judges, where the letters stand for positive and negative. The verdicts and types are labeled such that matching letters indicate a *correct* match, in the sense that the judge prefers these matches over their alternatives. The following table summarizes the phrases that can be used to refer to each verdict-type combination and clarifies how the two errors

(type-1 and -2) are defined.

Table 1: True/false positives and negatives

		Defendant Type (T)	
		P	N
Judge Decision (v)	p	True positive	False positive (Type-1 Error)
	n	False negative (Type-2 Error)	True negative

The proportion of type P people among the entire population of subjects is μ. However, with probability ρ, the judge observes and interprets a signal $S \in \{0, 1\}$. The signal is relevant and imperfectly informative of the subject's type, because $S = 1$ is more likely among type P subjects:

$$\phi_P \equiv P(S = 1|T = P) > \phi_N \equiv P(S = 1|T = N) \tag{1}$$

Thus, the judge finds herself in one of three states of the world, $s \in \{0, \varnothing, 1\}$, where 0 and 1 denote the nature of the signal when it is uncovered, and $\varnothing$ denotes the case where no signal is observed. The judge's beliefs about the proportion of type P subjects, given s, can be denoted as follows:

$$\gamma_s = \begin{cases} \mu & \text{if} \quad s = \varnothing \\ \frac{\mu(1-\phi_P)}{\mu(1-\phi_P)+(1-\mu)(1-\phi_N)} & \text{if} \quad s = 0 \\ \frac{\mu\phi_P}{\mu\phi_P+(1-\mu)\phi_N} & \text{if} \quad s = 1 \end{cases} \tag{2}$$

In each state s, the judge also reviews some evidence (θ), whose distribution $F(\theta|T)$ (with the associated probability density function $f(\theta|T)$) depends on the subject's type. Given this signal, the judge can select a threshold evidentiary requirement (in each state) which generates a type-1 error conditional on type and state, denoted $\alpha_s = P(v = p|T = N, s)$. As noted in the literature, when the evidentiary requirement is selected efficiently, the maximum correct positive probability, conditional on type, given any targeted α, can be expressed as an increasing and concave function $\beta(\alpha)$ with $\beta(0) = 0$ and $\beta(1) = 1$. I assume this function is twice differentiable, in which case $\beta'(\alpha) > 0 > \beta''(\alpha)$. Thus, $\beta_s = \beta(\alpha_s) = P(v = p|T = P, s)$ denotes the likelihood of a correct positive when the defendant is type P in state s.

A judge in state s chooses α_s to maximize her expected payoff, which has two components. The first is the difference between the expected benefit from a true positive relative to a false negative and is normalized to 1. The second is the the expected cost of a false positive relative to a true negative, denoted c. Thus, in each state the judge chooses α_s to maximize

$$\gamma_s\beta(\alpha_s) - (1 - \gamma_s)\alpha_s c \tag{3}$$

I assume that the signal and evidence are such that the judge chooses an interior α_s in all states,[2] in which case the judge's choices, α_s^*, are characterized by the following set of first order conditions:[3]

$$\beta'(\alpha_s^*) = \frac{1-\gamma_s}{\gamma_s} c \tag{4}$$

And, $\beta_s^* \equiv \beta(\alpha_s^*)$ denotes the probability of a correct positive in each state of the world.

It is worth noting that, the judge's decision making criteria characterized by (4) corrsponds to a threshold rule, where the judge returns a verdit of p when, conditional on the state of the world s that he observes, the probability of the subject being type P is greater than or equal to the critical threshold probability

$$\pi(c) = \frac{c}{1+c} \tag{5}$$

To see this note that (4) can alternatively be expressed as

$$P(T = P|\theta = \theta_s^*, s) = \frac{\gamma_s f(\theta_s^*|P)}{(1-\gamma_s) f(\theta_s^*|N) + \gamma_s f(\theta_s^*|P)} = \frac{c}{1+c} = \pi(c) \tag{6}$$

where θ_s^* is characterized by $F(\theta_s^*|N) = \alpha_s^*$, and thus the left-hand side of (6) corresponds to the probability with which the subject with the marginal characteristics, θ_s^*, is type P. The relationship expressed in (6) highlights how a judge's decision criterion relates to his preferences captured by c, and serves an important role in discussing taste-based discrimination in section 1.4.

For purposes of identifying monotonicity violations, it is also important to note that

$$\alpha_0^* < \alpha_\varnothing^* < \alpha_1^* \tag{7}$$

since $\phi_P > \phi_N$, and

$$\beta_0^* < \beta_\varnothing^* < \beta_1^* \tag{8}$$

because $\beta' > 0$.

Thus, the proportions of type N and type P subjects, respectively, who receive a verdict of p can be expressed as

$$R_N(\rho) \equiv \rho A + (1-\rho)\alpha_N^*; \text{ and} \tag{9}$$

$$R_P(\rho) \equiv \rho B + (1-\rho)\beta_N^* \tag{10}$$

where

$$A \equiv (1-\phi_N)\alpha_0^* + \phi_N \alpha_1^*, \text{ and} \tag{11}$$

$$B \equiv (1-\phi_P)\beta_0^* + \phi_P \beta_1^* \tag{12}$$

denote the proportions of type N and P subjects, respectively, receiving a p verdict, conditional on the observance of the signal S.

[2] For any c, this is true when the evidence generating process is sufficiently informative.

[3] Note that the concavity of β implies that the second order condition holds.

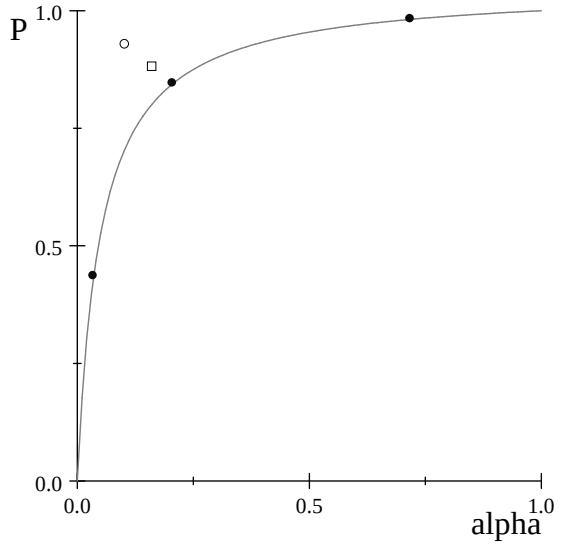
Using these observations, the true and false negative and positive rates can be defined and summarized by the following 'confusion matrix'

Table 2: Confusion Matrix (13)

		Defendant Type (T)	
		P	N
Judge Decision (v)	p	True positive rate $TP \equiv \mu R_P$	False positive rate $FP \equiv (1-\mu)R_N$
	n	False negative rate $FN \equiv \mu(1-R_P)$	True negative rate $TN \equiv (1-\mu)(1-R_N)$

Figure 1, below, depicts the concave β function, as well as the probability pairs $(\alpha^*_{s\in\{0,\varnothing,1\}}, \beta^*_{s\in\{0,\varnothing,1\}})$ (DOTS), A,B (CIRCLE), (R_P, R_N) (SQUARE), obtained through the functional forms in Example 1, below (in section 2.3.), that is used later in the analysis to provide a visual illustration of the potential relationship between these probabilities.

Figure 1



2.2 Heterogenous Judge Skill and Monotonicity (Violations)

The empirical literature exploiting random assignment of judges (discussed in more detail in subsection 3) typically relies on differences in judge *preferences* to account for differences in judge decision making, e.g., their propensities. This type of variation can be captured by differences in c, which reflects the relative

value they attach to avoiding one type of error rather than another. Here, instead, I focus on differences in judge skills, which, in turn, may affect the quality of their decisions. To isolate the impact of this consideration, I consider cases in which judges have a common c, but may differ in their ability to observe and interpret signals captured by differences in ρ.

An important question is whether judge decisions are *monotonic*, in the sense of the phrase used in the empirical literature, when there is skill heterogneity among judges. In this literature, monotonicity is a condition often assumed to ensure that estimated LATEs are valid. To relate judge behavior to this context, I define the following.

Definition 1 ***Probabilistic Monotonicity:*** *Judge behavior is monotonic if* $R_A(\rho') \geq R_A(\rho'')$ *and* $R_B(\rho') \geq R_B(\rho'')$ *or* $R_A(\rho') \leq R_A(\rho'')$ *and* $R_B(\rho') \leq R_B(\rho'')$ *for any* ρ', ρ''.

It is worth noting that this monotonicity definition is weaker than what may be termed *strict monotonicity* which would require monotonicity across the assignment of all individual defendants (i.e., a defendant who would be assigned to p by a judge with lower propensity would be assigned to p by a judge with greater propensity and vice versa). That type of montonicity would be trivially violated in this context, since $\alpha_0^* < \alpha_\varnothing^* < \alpha_1^*$. Nevertheless, probabilistic monotonicity is sufficient to recover valid LATEs in many circumstances, and thus I focus on this type of monotonicity here (I discuss other types of monotonicity and additional inference problems in section 4, below), and refer to it simply as monotonicity in this section.

With this definition is place, one can investigate whether judge behavior in this context is likely to be monotonic, and more generally, how R_A and R_B are likely to change with judge skills. The next proposition identifies a condition under which monotonicity is violated, and also provides a partial characterization of the relationship between judge skills and decisions as a function of the properties of the evidence generation process (henceforth 'EGP') summarized by the function β.

Proposition 1 *A. If the true positive rate is decreasing in judge skill, then so is the false positive rate.* $\frac{dR_B}{d\rho} \leq 0 \Longrightarrow \frac{dR_A}{d\rho} \leq 0$.

B. Let

$$C(\alpha) \equiv -\frac{\beta'(\alpha)}{\beta''(\alpha)} \tag{14}$$

(i) The true positive rate is increasing and the false positive rate is decreasing in judge skill (i.e., $\frac{dR_B}{d\rho} > 0 > \frac{dR_A}{d\rho}$*) if* $0 > C'(\alpha) > 1$ *for all* $\alpha \in [\alpha_0^*, \alpha_1^*]$*;*

(ii) Both true and false positive rates are increasing in judge skill (i.e., $\frac{dR_A}{d\rho}, \frac{dR_B}{d\rho} > 0$*) if* $C'(\alpha) > 1$ *for all* $\alpha \in [\alpha_0^*, \alpha_1^*]$*; and*

(iii) Both true and false positive rates are decreasing in judge skill (i.e., $\frac{dR_A}{d\rho}, \frac{dR_B}{d\rho} < 0$*) if* $C'(\alpha) < 0$ *for all* $\alpha \in [\alpha_0^*, \alpha_1^*]$.

Proof. Note that $\frac{dR_A}{d\rho} = A - \alpha_N^*$ and $\frac{dR_B}{da} = B - \beta_N^*$, and thus,

$$\begin{aligned} \frac{dR_A}{d\rho} &\leq 0 \text{ iff } \alpha_N^* - A \geq 0; \text{ and} \\ \frac{dR_B}{d\rho} &\leq 0 \text{ iff } \beta_N^* - B \geq 0. \end{aligned}$$

As noted in Lundberg and Mungan (2022), $\alpha_N^* - A > 0$ if $C'(\alpha) < 1$ (and $\alpha_N^* - A < 0$ if $C'(\alpha) > 1$) and $\beta_N^* - B > 0$ if $C'(\alpha) < 0$ (and $\beta_N^* - B < 0$ if $C'(\alpha) > 0$) for all $\alpha \in [\alpha_0^*, \alpha_1^*]$, which implies the statements in the proposition. ■

A natural implication of proposition 1 is that monotonicity can be violated in a variety of conditions. The next corollary summarizes this result.

Corollary 1 *Decisions made by judges with differing skills violate monotonicity when* $C'(\alpha) \in (0, 1)$ *for all* $\alpha \in [\alpha_0^*, \alpha_1^*]$.

Corollary 1 and Proposition 1 together note the technical properties that an EGP must possess for judges' behavior to be non-monotonic and monotonic in different directions as a function of judge skill. This raises the question of whether 'natural' evidence generating processes fall into one of the three technical categories identified in part B of the proposition. Although it is impossible to fully categorize EGPs into one of the three categories identified in part B of the proposition, there are some intuitive observations that can be made.

Remark 1 *(i) Suppose* $f(\theta|N)$ *and* $f(\theta|P)$ *are symmetric about some* $\hat{\theta}$ *and* $\frac{f(\theta|P)}{f(\theta|N)}$ *is monotonic. Then,* $C'(\bar{\alpha}) = 0.5$ *for* $\bar{\alpha}$ *such that* $\bar{\alpha} + \beta(\bar{\alpha}) = 1$. *Thus,* $0 > C'(\alpha) > 1$ *for all* $\alpha \in [\alpha_0, \alpha_1]$ *as long as* $|\alpha_0 - \bar{\alpha}|$ *and* $|\alpha_1 - \bar{\alpha}|$ *are not too large.*

(ii) Suppose $\beta(\alpha) = \frac{\alpha(1+k)}{\alpha+k}$ *for* $k > 0$. *Then,* $C'(\alpha) = 0.5$ *for all* α.

(iii) Suppose $\beta(\alpha) = \alpha^k$ *for some* $k \in (0, 1)$, *then* $C'(\alpha) = \frac{1}{1-k} > 1$.

(iv) Suppose $\beta(\alpha) = 1 - (1 - \alpha)^k$ *for some* $k > 1$, *then* $C'(\alpha) = \frac{1}{1-k} < 0$.

Part (i) of remark 1 notes sufficient conditions for $\frac{dR_B}{da} > 0 > \frac{dR_A}{da}$. As long as signals are generated in a symmetric manner, and α_s^* and $1 - \beta_s^*$ are not close to extremes, then the condition in part (i) of remark 1 holds. The symmetry and monotone likelihood ratio properties would hold, for instance, if the signal was generated through normal distributions that have equal variance, e.g., $F(\theta|T) = \Phi(\theta - \bar{\theta}_T)$ where Φ is the normal distribution with zero mean and variance of 1 and $\bar{\theta}_T$ denotes the mean evidence for type T. Parts (ii)-(iv), on the other hand, note specific families of functions for which conditions (i)-(iii) in proposition 2 hold, respectively, regardless of the specific values of α_0^* and α_1^*. Thus, all possibilities noted in the proposition can be obtained. However, it is much simpler to identify conditions (like symmetry) under which the result in part (i) of remark 2 holds.

Next, I consider the potential implications of monotonicity violations that occur due to variations in judge skill as opposed to judge preferences in a very simple setting.

2.3 Inference Problems: Monotonicity Violations and LATEs

I consider a researcher who attempts to identify the effect of verdicts on an outcome of interest denoted Y, e.g., the marginal impact of a conviction ($v = p$ rather than $v = n$) on the likelihood of future criminal behavior ($Y \in [0, 1]$). The true, and unobservable effect of a verdict on Y depends on the subjects type, i.e., $y = y(v, T)$. In this section, I abstract from the effect other unobservables and observables to clearly illustrate the inference problems that can be caused due to judge skill induced monotonicity violations. (In section 4, I explain how the presence of additional factors unobservable to the researcher, but observable to the judge, cause complications in designing tests and identification of treatment effects). Thus, the type-dependent impacts of a verdict on outcomes can be expressed as

$$\Delta_T \equiv y(p, T) - y(n, T) \text{ for } T \in \{P, N\}$$

and are unobservable. The researcher may be interested in estimating these effects (i.e., Δ_P and Δ_N), or the average impact of a verdict of p rather than n within a certain group. In the latter case, the effect would be given by a weighted average of the to average treatment effects by type. However, because Δ_P and Δ_N are unobservable, the researcher cannot directly identify (local) average treatment effects through a simple OLS regression, because v is correlated with T which is in turn correlated with y, which causes an endogeneity problem. To address this problem, a researcher may instead use an instrumental variables (IV) approach, wherein the propensity of a judge to return a p-verdict is used as an instrument to assess how verdicts affect outcomes. In this set-up, the propensity of a judge with skill ρ is given by

$$\Psi(\rho) = \mu R_B(\rho) + (1 - \mu) R_A(\rho)$$

and the outcomes associated with a judge of skill ρ is given by

$$Y(\rho) = (1 - \mu)[y(l, N) + R_A(\rho)\Delta_N] + \mu[y(l, P) + R_B(\rho)\Delta_P]$$

Given two judges with different skills ρ'' and ρ', the local average treatment effect (LATE) could be estimated by the ratio between the difference in judges' outcomes (i.e., $Y(\rho'') - Y(\rho')$) and propensities (i.e., $\Psi(\rho'') - \Psi(\rho')$). It is well known in the literature that this approach may yield biased results when monotonicity is violated. Here, I show that when judges act non-monotonically, the estimated LATEs obtained through this approach may not only be biased, but may carry the incorrect sign. Although this possibility is considered in other settings (Angrist et al. 1996), it can occur in a variety of circumstances when monotonicity is violated due to quality variations across judges.

To demonstrate this, note that using the approach described above yields the following LATE estimate:

$$\lambda(\rho', \rho'') = \frac{\Delta_Y(\rho', \rho'')}{\Delta_\Psi(\rho', \rho'')} \tag{15}$$

where

$$\begin{aligned} \Delta_Y(\rho', \rho'') &= Y(\rho') - Y(\rho''); \text{ and} \\ \Delta_\Psi(\rho', \rho'') &= \Psi(\rho') - \Psi(\rho'') \end{aligned} \tag{16}$$

It is worth noting that in an actual IV design, one could use many different judges with different propensities to estimate LATEs, and thus the estimated LATE would not be a function of only the two judges behavior and outcomes as in (15). Nevertheless, I focus on this case due to two important reasons. First, it illustrates the problems that can be generated due to monotonicity violations in a straightforward manner. Second, and more importantly, the tests I propose in section 4 to identify (the signs of) average treatment effects use the types of pairwise comparisons captured by (15).

I distinguish λ, the LATE estimate that the researcher would obtain through an IV design, from the true (and unobservable) LATE, which expresses the average treatment effect among the marginal population, i.e., people of the same type who have nevertheless been assigned different verdicts by the two judges. I denote the latter L, which can be expressed as

$$L(\rho', \rho'') \equiv \frac{\mu|R_B(\rho') - R_B(\rho'')|\Delta_P + (1-\mu)|R_A(\rho') - R_A(\rho'')|\Delta_N}{\mu|R_B(\rho') - R_B(\rho'')| + (1-\mu)|R_A(\rho') - R_A(\rho'')|} \tag{17}$$

The only difference between λ and L is that the former uses the actual differences between the positive verdict rates of the two judges whereas the latter uses the magnitudes of these difference. Naturally, under monotonicity the two measures are one and the same, as noted by the following.

Remark 2 *$\lambda(\rho', \rho'') = L(\rho', \rho'')$ for all ρ', ρ'' if judge decisions are monotonic.*

On the other hand, when judge decisions are not monotonic, λ and L may only coincidentally equal each other, and in general may differ from each other in many ways. Here, I identify the conditions under which the two measures may actually have the opposite sign, policy recommendations based on biased LATEs will be particularly misleading in such cases. Without loss of generality, I focus on the case where $\lambda > 0 > L$, and conditions can be identified for the opposite case using similar steps.

Proposition 2 *A) Monotonicity violations can cause the estimated LATE to have the wrong sign (i.e., it is possible that $L(\rho', \rho'')\lambda(\rho', \rho'') < 0$ is possible).*

B) Suppose the decisions of judges with skills ρ' and ρ'' violate monotonicity, and $\Psi(\rho') > \Psi(\rho'')$. (i) If $R_B(\rho') < R_B(\rho'')$, then $\lambda > 0 > L$ if, and only if,

$$-\Delta_P > \frac{(1-\mu)(R_A(\rho') - R_A(\rho''))}{\mu(R_B(\rho'') - R_B(\rho'))}|\Delta_N| \tag{18}$$

which can hold only if $\Delta_P < 0$, and $|\Delta_N|$ is not large.

(ii) If $R_B(\rho') > R_B(\rho'')$, then $\lambda > 0 > L$ if, and only if,

$$-\Delta_N > \frac{\mu(R_B(\rho') - R_B(\rho''))}{(1-\mu)(R_A(\rho'') - R_A(\rho'))}|\Delta_P| \tag{19}$$

which can hold only if $\Delta_N < 0$ *and* $|\Delta_P|$ *is not large.*

Proof. A) See example 1, below.

B) In general, $L < 0$ is equivalent to

$$\frac{(1-\mu)|R_A(\rho') - R_A(\rho'')|}{\mu|R_B(\rho') - R_B(\rho'')|}\Delta_N < -\Delta_P \tag{20}$$

and $\lambda > 0$ is equivalent to

$$(1-\mu)(R_A(\rho') - R_A(\rho''))\Delta_N > -\Delta_P\mu(R_B(\rho') - R_B(\rho'')) \tag{21}$$

(i) If $R_B(\rho') < R_B(\rho'')$, then (20) can be written as

$$-\Delta_P > \Delta_N\frac{(1-\mu)(R_A(\rho') - R_A(\rho'')}{\mu(R_B(\rho'') - R_B(\rho'))})$$

and (21) can be written as

$$-\Delta_P > -\Delta_N\frac{(1-\mu)(R_A(\rho') - R_A(\rho'')}{\mu(R_B(\rho'') - R_B(\rho'))})$$

Thus, both (20) and (21) holds if, and only if (18) holds.

(ii) If $R_B(\rho') > R_B(\rho'')$, then (20) can be written as

$$-\Delta_N > \Delta_P\frac{\mu(R_B(\rho') - R_B(\rho''))}{(1-\mu)(R_A(\rho'') - R_A(\rho'))}$$

and (21) can be written as

$$-\Delta_N > -\Delta_P\frac{\mu(R_B(\rho') - R_B(\rho''))}{(1-\mu)(R_A(\rho'') - R_A(\rho'))}$$

Thus, both (20) and (21) holds if, and only if (19) holds. ■

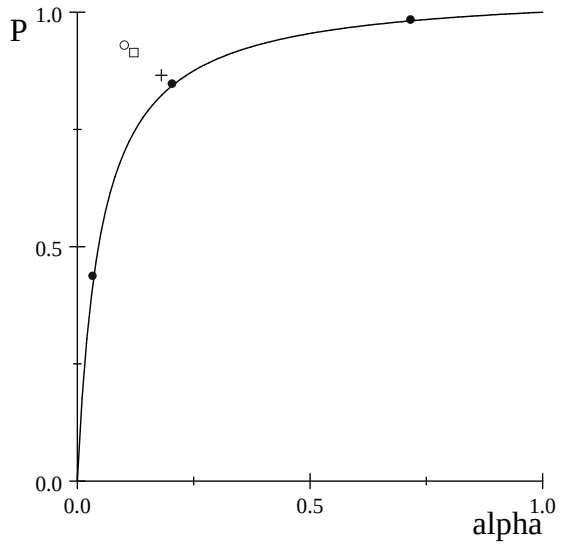
Proposition 2 notes that for the estimated LATE to be negative while the true LATE is positive, the treatment effect for the type for which the judge with the higher propensity returns positive verdicts more frequently must be negative. That the conditions specified in proposition 2 can hold is proven next through an illustrative example.

Example 1 $\beta(\alpha) = \frac{\alpha(1.05)}{\alpha+0.05}$ $\rho' = 1/4$, $\rho'' = 4/5$, *and* $\Delta_P = -0.1$, $\Delta_N = 0.1$, $y(l, L) = 0$, $y(l, H) = 0.7$, $\phi_P = 0.9$, $\phi_N = 0.1$, $c = 0.8$, $\mu = 0.5$.

The β *curve, along with the probabilities of* $\alpha^*_{i\in\{0,\varnothing,1\}}$, $\beta^*_{i\in\{0,\varnothing,1\}}$, A, B, *and* $R_{i\in\{A,B\}}$ *for the two judges in this example can be depicted, as follows. Here, the dots represent* α^*_i, β^*_i *pairs, the circle represents the* A, B *pair, the cross represents the* $R_A(\rho'), R_B(\rho')$ *pair, and the box represents the* $R_A(\rho''), R_B(\rho'')$

pair.

Figure 2



The choices and characteristics of the two judges, along with the true and estimated LATEs can be summarized as follows (rounded off to the nearest 100th):

$$\begin{array}{cccc} & \rho' = 1/4 & & \rho'' = 4/5 \\ \Psi & 52.3\,1\% & > & 51.7\,4\% \\ Y & 26.53\% & > & 26.02\% \end{array}$$

$$\lambda(\tfrac{1}{4}, \tfrac{4}{5}) = 0.9; \quad L(\tfrac{1}{4}, \tfrac{4}{5}) = -0.08$$

This demonstrates that a small and negative effect on the outcome of interest can appear as if it is large and positive. Hence, monotonicity violations can cause the estimated LATE to not only be biased, but for it to have the wrong sign.

2.4 Inference Problems: Judge Decision Criteria and Discrimination

Here, I explain additional complications caused by the presence of heterogeneity in judge skills. Specifically, when judges differ in quality as explained here, the average outcome rates of marginal subjects cannot be used as a proxy of the decision criterion used by judges. This is a strategy adopted in Arnold et al. (2018) in the pretrial release context. The setting they study is one in which a judge decides whether or not to (continue to) detain or release a defendant prior to their trial. The data they analyze allow them to observe whether released defendants commit acts which qualify as a 'pretrial misconduct', e.g., failure to

appear in court, or the commission of a new crime. In this setting, the authors estimate the average pretrial misconduct rates of marginal defendants, and argue that these can be used as estimates of the decision criteria used by judges to determine whether to release a defendant (e.g., release if likelihood of pretrial misconduct is greater than x). The authors then compare these estimates across defendants of different races. Based on these findings, they suggest that judges must be engaging in discrimination. Canay et al. (2022) identify several problems with this approach, including the invalidity of the outcomes test adopted in the article. The problems I identify here complement those identified by Canay et al. and do focuses on why the estimated probabilities cannot serve as proxies of judge decision criteria.

The set-up described in the previous sections accomodate the pretrial misconduct setting studied in the literature. Specifically, by letting 'positives' refer to releases (i.e., $v = p$) and individuals who do not commit pretrial misconducts when released (i.e. $T = P$) one can reinterpret the setup previously described (e.g., a false position where $v = p$ and $T = N$ corresponds to the releasing of a person who subsequently commits a pretrial misconduct).

2.4.1 Outcomes and Propensities

In this setting, the judge perceives a subject belonging to one of three populations $s \in \{\varnothing, 0, 1\}$) with three probabilities: ρ; $(1-\rho)\zeta_0$ and $(1-\rho)\zeta_1$, where $\zeta_{i\in\{0,1\}} = P(S = i)$. The outcome variable of interest, Y, is the pretrial misconduct rate. Note that given the decision criterion adopted by the judge, the pretrial misconduct rates within each population ($\varnothing, 1$, and 0) can be expressed as $Y^{\varnothing}$, Y^0 and Y^1. Thus, the pretrial misconduct rates associated with a judge of skill ρ is given by:

$$Y(\rho) = (1-\rho)Y^{\varnothing} + \rho(\zeta_0 Y^0 + \zeta_1 Y^1) \tag{22}$$

Similarly, if we denote the rate of release (Ψ) for each of these three categories as Ψ_G, Ψ_0 and Ψ_1, then it follows that the judge's release propensity is given by:

$$\Psi(\rho) = (1-\rho)\Psi^{\varnothing} + \rho[\zeta_0 \Psi^0 + \zeta_1 \Psi^1] \tag{23}$$

2.4.2 Two Judges, Single Race: Inferring Judge Decision Criteria?

Next, consider two judges who only differ in their skill, ρ, and with skill levels ρ_1 and ρ_2.The average pretrial misconduct rates for the marginal defendants for these two judges will be given by:

$$\begin{aligned}
\lambda(\rho_1, \rho_2) &= \frac{Y(\rho_1) - Y(\rho_2)}{\Psi(\rho_1) - \Psi(\rho_2)} = \frac{(\rho_1 - \rho_2)(Y^{\varnothing} - \zeta_0 Y^0 - \zeta_1 Y^1)}{(\rho_1 - \rho_2)(\Psi^{\varnothing} - \zeta_0 \Psi^0 - \zeta_1 \Psi^1)} \\
&= \frac{Y^{\varnothing} - \zeta_0 Y^0 - \zeta_1 Y^1}{\Psi^{\varnothing} - \zeta_0 \Psi^0 - \zeta_1 \Psi^1} = \bar{\lambda}
\end{aligned}$$

Note that $\bar{\lambda}$ is constant, i.e., it does not depend on the skill level. Moreover, this value, in general, does not equal $\pi(c) = \frac{c}{1+c}$ (as defined in (5)), the critical probability of pretrial misconduct among marginal offenders. This is because $\bar{\lambda}$ is not directly related to the judges decision criterion on the margin. Instead, it captures the ratio between the reduction in false positives (incorrect releases) and the reduction in overall releases caused by the acquisition of superior information. The two concepts are generally unrelated. This can be demonstrated further by expressing the pretrial misconduct rates as a function of type-1 errors (i.e. α) and priors (i.e. μ) and similarly expressing propensities as a function of release rates (i.e. α and β) as well as priors, which reveals that

$$\bar{\lambda} = \frac{(1-\mu)(\alpha^*_\varnothing - A)}{\mu(\beta^*_\varnothing - B) + (1-\mu)(\alpha^*_\varnothing - A)} \tag{24}$$

The expression in (24) reveals that a variant of the *inframarginality problem* resurfaces when judges differ in quality: $\bar{\lambda}$ is not a function only of the characteristics of the defendants that the judge decides to release on the margin, but also of the characteristics of the inframarginal defendants (since $\alpha^*_\varnothing - A$ as well as $\beta^*_\varnothing - B$ are affected by the characteristics of inframarginal defendants). Thus, there is no reason to expect $\bar{\lambda}$ to track π, and this can be verified through simple examples.

2.4.3 Inferring Discrimination

That $\bar{\lambda} \neq \pi$ implies that one cannot use estimates of $\bar{\lambda}$ to conduct outcome tests to identify discrimination. Moreover, differences in $\bar{\lambda}$ are quite natural. To see this, note that the above steps can be repeated for two different races two obtain two separate $\bar{\lambda}$'s, one for each race. Differences in races can be captured, among other things, through the proportion of type P defendants awaiting release, i.e., μ, which may differ across races due to many non-judge discrimination related reasons (e.g., discrimination at the policing stage). In general $\frac{d\lambda}{d\mu} \neq 0$, which implies that differences in $\bar{\lambda}$ are natural.

2.5 Additional Problems when Decision Quality is Endogenous

Thus far, the analysis focused on variations in judge decision quality (captured by variations in ρ) which are exogenously determined. Here, I consider a case where ρ emerges as a result of costly efforts (denoted e) by judges (e.g., information analysis) to increase the quality of their decisions such that $\rho = \rho(e)$ with $\rho' > 0$. Here, e denotes the type of effort or analysis required to generate a probability of ρ. To capture skill variation across judges, I assume that the cost to judges from conducting this type of analysis equals e/a, such that a higher a corresponds to greater skill.

Thus, a judge chooses e to maximize her expected payoff which can be expressed as:

$$\Pi = \sum_s q_s(e)(\gamma_s \beta_s^* - (1-\gamma_s)\alpha_s^* c) - \frac{e}{a} \tag{25}$$

where

$$q_s(e) = \begin{cases} 1-\rho(e) & \text{if} \quad s = \varnothing \\ \rho(e)P(S=0) & \text{if} \quad s = 0 \\ \rho(e)P(S=1) & \text{if} \quad s = 1 \end{cases} \tag{26}$$

and it is useful to note that

$$\begin{aligned} P(S &= 0) = \mu(1-\phi_P) + (1-\mu)(1-\phi_N); \text{ and} \\ P(S &= 1) = \mu\phi_P + (1-\mu)\phi_N \end{aligned} \tag{27}$$

The investment of a judge e^*, is thus characterized by the first order condition

$$\Pi_e = \sum_s q_s'(e^*)(\gamma_s \beta_s^* - (1-\gamma_s)\alpha_s^* c) - \frac{1}{a} = 0 \tag{28}$$

and depends on his skill level a, i.e., $e^* = e^*(a)$.

2.5.1 The Effect of c on Outcomes

A change in c affects both decisions of a judge, i.e., α_s^*'s as well as e^*. The next proposition summarizes these impacts, as follows.

Remark 3 *Increasing the cost of false negatives (i) reduces all state dependent probabilities of positive verdicts, i.e.,* $\frac{d\alpha_s^*}{dc}, \frac{d\beta(\alpha_s^*)}{dc} < 0$*, and (ii) reduces (resp. increases) the quality of decisions if* $A - \alpha_\varnothing^* < 0$ *(resp.* > 0*), i.e., if the observation of additional evidence reduces false negatives in expectation.*

Proof. (i) From (4) it follows that $\frac{d\alpha_s^*}{dc} = \frac{\frac{1-\gamma_s}{\gamma_s}}{\beta''(\alpha_s^*)} < 0$. (ii) From (28) it follows that $\frac{de^*}{dc} = -\frac{\rho'(e^*)(1-\mu)(A-\alpha_\varnothing^*)}{\rho''(e^*)(\mu[B-\beta_\varnothing^*]-(1-\mu)c[A-\alpha_\varnothing^*])}$, and thus, $sign(\frac{de^*}{dc}) = sign(A - \alpha_\varnothing^*)$. ■

Although the impact of an increase in the cost of false negatives on state-dependent verdict probabilities and the quality of decisions is relatively straightforward, its impact on the expected probabilities of verdicts, by type, is not as clear. These impacts can be ascertained as follows

$$\frac{dR_A}{dc} = \overbrace{\underbrace{\rho(e^*)\frac{dA}{dc} + (1-\rho(e^*))\frac{d\alpha_\varnothing^*}{dc}}_{\textit{Effect through smaller } \alpha_s^*}}^{(-)} + \overbrace{\underbrace{\rho'(e^*)(A-\alpha_\varnothing^*)\frac{de^*}{dc}}_{\substack{\textit{Effect through change} \\ \textit{in decision quality}}}}^{(+)} \tag{29}$$

Thus, the impact of a change in c on false positives is, *a priori*, unclear, and depends on the comparison between the effect identified in remark 3-(i) and the

impact due to the change in judge decision quality. The former effect is negative, while the latter is positive. Thus, the overall impact cannot be ascertained absent further analysis.

The impacts of c on the likelihood of correct positives, on the other hand, are even more complex, which is revealed through the following:

$$\frac{dR_B}{dc} = \overbrace{\underbrace{\rho(e^*)\frac{dB}{dc} + (1-\rho(e^*))\frac{d\beta(\alpha^*_\varnothing)}{dc}}_{\textit{Effect through smaller } \beta(\alpha^*_s)}}^{(-)} + \overbrace{\underbrace{\rho'(e^*)(B-\beta(\alpha^*_\varnothing))\frac{de^*}{dc}}_{\textit{Effect through change in decision quality}}}^{(+) \text{ iff monotonic}} \quad (30)$$

An interesting implication of this observation is that, when monotonicity is violated, the probability of correct convictions is reduced. This is summarized by the following, and its implication is noted, as follows:

Proposition 3 *(i) Increasing the cost of false negatives reduces the true positive rate when there is a violation of monotonicity. (ii) Thus, if 'miss rates' are defined as*

$$m(c) \equiv \mu(1 - R_B(c))$$

then $\frac{dm}{dc} < 0$ *implies that judge behavior is monotonic.*

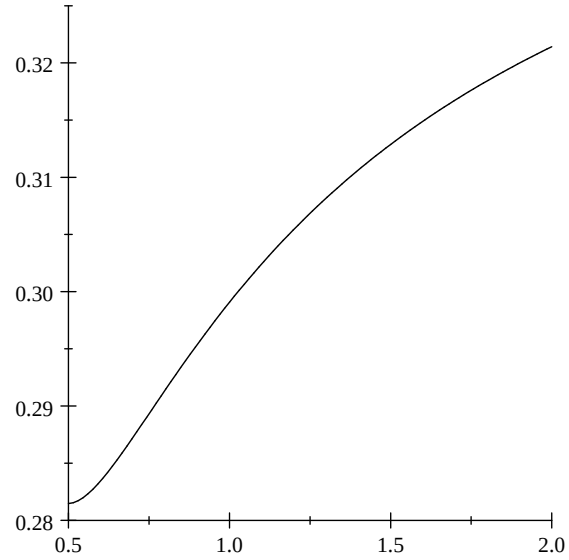
Proof. *(i) Follows directly from (30). (ii)* $\frac{dm}{dc} = -\mu\frac{dR_B}{dc} < 0 \implies \frac{dR_B}{dc} > 0 \implies sign(B - \beta(\alpha^*_\varnothing)) = sign(A - \alpha^*_\varnothing)$. ■

Next, I consider a situation where monotonicity is not violated and the researcher finds a positive and unbiased LATE, where a positive outcome is unfavorable (e.g., recidivism). Based on this finding, the researcher suggests that increasing c can reduce judge's inclination's to return positive verdicts, and thereby reduce the outcome variable. I question whether this policy implication can be misleading through an illustrative example where the two average treatment effects by type have opposite signs.

Example 2 $\beta(\alpha) = \sqrt{\alpha}$ *and* $p(e) = e(2-e)$ *with* $\bar{e} = 1$*, and* $\Delta_P = -0.2$*,* $\Delta_N = 0.1$*,* $y(l, L) = 0$*,* $y(l, H) = 0.7$*,* $\phi_P = 0.7$*,* $\phi_L = 0.4$*,* $\mu = 0.5$*.*

The average outcomes (i.e., Y*) for a judge with skill* $a = 50$ *around as a function of* c *around* $c = 1$ *can then be depicted, via figure 3, as follows.*

Figure 3



Example 2 illustrates the possibility of misleading policy implications based on valid LATEs when the two average treatment effects differ, and in particular, when they have opposite signs. This naturally raises the question of whether there are any plausible mechanisms which may cause the treatment effects to differ from each other, and perhaps cause them to have opposite signs. Next, I consider a simple model that provides a rationale for this possibility.

3 Opposite Signed Treatment Effects by Type: A Simple Law Enforcement Model

I consider a setting similar to that of Miceli et al. (2022), but which incorporates the possibility of erroneous punishment. A person decides whether to commit crime in each period in a two period setting. His benefit from crime in each period is drawn from the distribution $G(b)$. The magnitude of the sanction is normalized to 1, is the same across the two periods, and is common knowledge. On the other hand, the probability of punishment, which depends on whether the person actually commits crime, is unobservable. The person believes that if he does not commit crime, the true probability of (wrongful) punishment is $\varepsilon \in \{\varepsilon_l, \varepsilon_h\}$ with $\varepsilon_h > \varepsilon_l$. If he commits crime, he believes the probability of punishment is $p \in \{p_l, p_h\}$ with $p_h > p_l$. His prior beliefs are described by $q_0 = P(\varepsilon = \varepsilon_l)$ and $q_1 = P(p = p_l)$, and he also believes that these two probabilities are independently determined. Thus, ex-ante, the person believes

that the expected probability of punishment is $\hat{\varepsilon} = q_0\varepsilon_l + (1-q_0)\varepsilon_h$ if he refrains from committing crime, and $\hat{p} = q_1 p_l + (1-q_1)p_h$ if he commits crime.

To analyze this person's decision making process, first, note that he enters the second period in one of four states, summarized by the following table, which lists the notation I use to describe each state $\sigma \in \{TP, TN, FP, FN\}$.

$$\begin{array}{lll} & \text{Commit Crime} & \text{Don't} \\ \text{Punished} & TP & FP \\ \text{Not} & FN & TN \end{array} \tag{31}$$

The person's beliefs under these four states of the world about the two probabilities of detection, listed in the form (p, q), are thus given by the following:

$$\begin{array}{lll} & \text{Commit Crime} & \text{Don't} \\ \text{Punished} & (\frac{q_1 p_l^2 + (1-q_1)p_h^2}{q_1 p_l + (1-q_1)p_h}, \hat{q}) & (\hat{p}, \frac{q_0\varepsilon_l^2 + (1-q_0)\varepsilon_h^2}{q_0\varepsilon_l + (1-q_0)\varepsilon_h}) \\ \text{Not} & (\frac{p_l[1-p_l]q_1 + p_h[1-p_h](1-q_1)}{1-q_1 p_l - (1-q_1)p_h}, \hat{q}) & (\hat{p}, \frac{\varepsilon_l[1-\varepsilon_l]q_0 + \varepsilon_h[1-\varepsilon_h](1-q_0)}{1-q_0\varepsilon_l - (1-q_0)\varepsilon_h}) \end{array} \tag{32}$$

Therefore, if the person was punished after committing crime in the first period, he will commit crime in the second period if:

$$b > b_{TP} \equiv \frac{q_1 p_l^2 + (1-q_1)p_h^2}{q_1 p_l + (1-q_1)p_h} - \hat{q} \tag{33}$$

On the other hand, if he avoided punishment despite committing crime in the first period, he will commit crime in the second period if:

$$b > b_{FN} \equiv \frac{p_l[1-p_l]q_1 + p_h[1-p_h](1-q_1)}{1-q_1 p_l - (1-q_1)p_h} - \hat{q} \tag{34}$$

Thus, the impact of punishment on the likelihood of the person committing crime, if he was truly guilty in the first period, is given by

$$\Delta_P = G(b_{FN}) - G(b_{TP}) \tag{35}$$

Note that $\Delta_P < 0$ as long as $p_h < p_l$, thus the treatment effect when the person is truly guilty is negative, i.e., punishment reduces the likelihood of recidivism.

On the other hand, if the person does not commit crime in the first period but is nevertheless punished, he commits crime in the second period if:

$$b > b_{FP} \equiv \hat{p} - \frac{q_0\varepsilon_l^2 + (1-q_0)\varepsilon_h^2}{q_0\varepsilon_l + (1-q_0)\varepsilon_h}$$

Similarly, if the person refrains from committing crime and is not punished in the first period, he commits crime in the second period if:

$$b > b_{TN} \equiv \hat{p} - \frac{q_0\varepsilon_l^2 + (1-q_0)\varepsilon_h^2}{q_0\varepsilon_l + (1-q_0)\varepsilon_h} \tag{36}$$

Thus, the impact of punishment on the likelihood of the person committing crime, if he was truly innocent in the first period, is given by

$$\Delta_N = G(b_{TN}) - G(b_{FP}) \tag{37}$$

which is positive as as long as $\varepsilon_h > \varepsilon_l$. Therefore, the treatment effect associated with the punishment of an innocent individual is positive, i.e., punishment increases the likelihood of committing crime.

The model presented here illustrates the learning dynamics that can cause treatment effects to differ by type (i.e., innocent or guilty). It is not meant to suggest that these are the only dynamics that affect treatment effects. A very large body of work investigates the likely effects of punishment on people's propensities to recidivate, and identifies many competing considerations that may reduce or increase people's propensities. The objective is instead to note that these effects can differ by type, and thus they may plausibly have opposite signs. As I noted in section 2, this can cause LATEs estimated through IV designs to be invalid, and may also cause policy implications based on valid LATEs to be misleading. Thus, I investigate whether average treatment effects by type can be identified even when monotonicity conditions are violated.

4 Identification of Treatment Effects with Unobservables

In section 2, I abstracted from all considerations that may affect treatment effects as well as judges' decisions, with the exception of the evidence and the signal received by judges. I did this to isolate the impact of endogenously determined judge decisions on monotonicity, and to focus on inference problems that may emerge in IV designs exploiting variation in judge propensities. However, in reality, there can be additional complications caused by the presence of factors that are unobservable to the researcher, but which are observable to the judge. Additional inference problems can arise when these factors affect judge decisions or treatment effects. Here, I incorporate these factors, to formalize these inference problems, and to question under what circumstances treatment effects can be validly estimated.

For this purpose, I refer to all factors unobservable by the researcher but observable by the judge simply as *unobservables*. I distinguish between two types of unobservables. First, the signal (θ) whose generation is affected by the subject's type T, and all other unobservables $u \in U$. I allow for inferences made by the judge based on θ to be affected by u as well as the judge's own characteristics, and thus express the probabilities with which a type N and type P subject, respectively, receive a verdict of p conditional on being assigned to judge i as $\hat{\beta}^i(u)$ and $\hat{\alpha}^i(u)$. I no longer constrain these probabilities to be obtained through the first order conditions noted in (4) in section 2.1, and instead allow them to be formed based on an any decision criteria a judge may adopt. The signal θ is what allows the judge to return a p verdict with different

probabilities to people of different types but possessing the same u, and the potential dependency of $\hat{\beta}^i(u)$ and $\hat{\alpha}^i(u)$ on u reflects that the signal θ is allowed to interact with other unobservables. To illustrate this, consider the specific case analyzed in section 2 where the only additional unobservable is the signal state $s \in \{0, \varnothing, 1\}$ such that $\hat{\beta}^i(1) > \hat{\beta}^i(\varnothing) > \hat{\beta}^i(0)$ and $\hat{\alpha}^i(1) > \hat{\alpha}^i(\varnothing) > \hat{\alpha}^i(0)$. It is also worth noting that some unobservables may also relate to the characteristics of subjects, which, in turn, may have an impact on treatment effects, which I denote

$$\hat{\Delta}_T(u) \text{ for } T \in \{P, N\} \tag{38}$$

Thus, due to the presence of unobservables, it is no longer possible to define monotonicity as in definition 1, which only considers two kinds of subjects distinguished only by their types. When unobservables are present, a more demanding type of probabilistic monotonicity requirement for validity can be defined as follows.

Definition 2 *Global monotonicity (GM): For any judge pair i, j: Either $\hat{\beta}^i(u) \geq \hat{\beta}^j(u)$ and $\hat{\alpha}^i(u) \geq \hat{\alpha}^j(u)$ for all u, or $\hat{\beta}^i(u) \leq \hat{\beta}^j(u)$ and $\hat{\alpha}^i(u) \leq \hat{\alpha}^j(u)$ for all u.*

It is worth noting that this monotonicity requirement is of the weaker, probabilistic, kind. However, it is global, in the sense that it requires a ranking across all types and all other unobservables. Thus, it is stricter than a monotonicity requirement that has been tested in the literature, which focuses on monotonicity based on verdict rates by type (e.g., Chan et al. (2022) and Bhuller and Sigstad 2022). To define this second conception of monotonicity, one can define average verdict rates, as follows:

$$\beta^i \equiv \int_{u \in U} \hat{\beta}^i(u) dG(u); \text{ and } \alpha^i \equiv \int_{u \in U} \hat{\alpha}^i(u) dG(u) \tag{39}$$

where $G(u)$ is the distribution of u. Using these averages, one can define a weaker monotonicity requirement, as follows.

Definition 3 *Across-type monotonicity (ATM): $\beta^i \geq \beta^j \iff \alpha^i \geq \alpha^j$ for all i, j.*

It is important to note that GM is sufficient to recover valid LATEs in this set-up (where exclusivity and existence are not relevant), but ATM is not. I note the former result, and prove the second one, as follows.

Proposition 4 *(i) GM implies that estimated LATEs are unbiased. (ii) Estimated LATEs can be biased even when ATM holds.*

Proof. (i) The generalized version of this result is proven in Chan et al. (2022) who show that GM (which they call probabilistic monotonicity) implies what Frandsen et al. (2019) call 'average' monotonicity, which Frandsen et al. (2019) show implies that estimated LATEs are unbiased.

(ii) Example 3, in section 4.2, below, represents a case where ATM holds and the estimated LATE is nevertheless biased, and thus constitutes a proof. ■

Proposition 4 highlights the importance of distinguishing between global monotonicity and the weaker concept of across-type monotonicity. Although global monotonicity, when combined with the standard requirements of exclusivity and existence yields valid LATE estimates, the weaker condition of across-type monotonicity does not. Thus, although violations of ATM are sufficient to suggest that LATE estimates are invalid (as in Chan et al. (2022)), ruling out violations of ATM (as in Bhuller and Sigstad 2022), alone, should not provide confidence that LATE estimates are valid.

This raises the question of whether there are conditions under which the two types of monotonicity are equivalent. Since the difference between the two types of monotonicity are driven by the impact of unobservables not exclusively related to type on verdicts, it naturally follows that when judges ignore (or are unable to detect) these unobservables the two types of monotonicity are equivalent. To formalize this result, I first define this condition, as follows.

Definition 4 *Decision irrelevance of other unobservables (DIU): For all i there exists $\bar{\beta}^i$ and $\bar{\alpha}^i$, such that $\hat{\beta}^i(u) = \bar{\beta}^i$ and $\hat{\alpha}^i(u) = \bar{\alpha}^i$ for all u.*

The condition defined may require clarification and interpretation. When DIU holds, there is a single unobservable which affects the decisions of judges, and it is a signal that is related only to the type (i.e., $T \in \{P, N\}$) of the subject. No other unobservables affect the decision process of judges. The condition appears strong, but may be reasonable in some cases, e.g., when the classifier is a radiologist attempting to identify pneumonia by studying X-Rays (as in Chan et al. (2022)). Moreover, it can also be *required* to hold by law in some cases where a decision maker has to make a determination based on a single input, e.g., a police officer determining whether to issue a speeding ticket based on the reading on his speed detector. It is important to note that DIU does not imply that there is no selection on unobservables: A judge still observes a signal, θ, which is not observed by the researcher, which affects what type of verdict she returns. DIU also allows for across-judge variation in decisions. Different judges may interpret θ differently, e.g., when they have exogenously determined and differing decision qualities, and thus may therefore have different verdict rates.

The next proposition highlights the relationship between the two types of monotonicity and DIU.

Proposition 5 *ATM and GM are equivalent if, and only if, DIU holds.*

Proof. GM$\Longrightarrow$ATM regardless of whether DIU holds. Thus, whether ATM and GM are equivalent depend on whether ATM$\Longrightarrow$GM. If DIU holds, then $\hat{\beta}^i(u) = \beta^i$ and $\hat{\alpha}^i(u) = \alpha^i$ for all i, and therefore GM$\Longrightarrow$ATM. Example 3 in section 4.2, below, represents a case where ATM holds but DIU and GM are violated. ■

A natural implication of proposition 5 is that the weaker concept of ATM can be used in place of GM, when DIU holds, which is noted by the following corollary.

Corollary 2 *ATM combined wirth DIU implies that LATE estimates are valid.*

Although ATM combined with DIU implies that estimated LATEs will be valid, DIU may appear to be a condition that is too strong. For instance, in the endogenous judge decision making model in section 2, if the states of the world $s \in \{0, \varnothing, 1\}$ are observed with different frequencies based on an unobservable characteristic of the subject that is not exclusively a function of his type, then DIU will be violated. It is therefore important to note that although DIU is sufficient to make LATE estimates valid when ATM holds, it is not necessary. The reason that DIU allows inferences when ATM holds is that the outcome differential across two judges 1, 2, i.e., $\Delta_Y^{1,2}$, are then a function only of the average verdict rates by type (i.e., β^i and α^i for $i \in \{1, 2\}$) and the average treatment effect by type, which can be defined as:

$$\Delta_T \equiv \int_U \hat{\Delta}_T(u) dG(u) \text{ for } T \in \{P, N\}$$

Thus, a weaker sufficient condition than DIU, which when combined with ATM implies valid LATEs is the following.

Condition 1 $\Delta_Y^{i,j} = \mu(\beta^i - \beta^j)\Delta_P + (1-\mu)(\alpha^i - \alpha^j)\Delta_N$ *for all* i, j.

Condition 1 requires the outcome differential between any two judges to be given by the sum across the two types of the expected positive decision rate differential by type times the average treatment effect by type. Next, I note two additional properties (besides DIU) which imply condition 1. The second of these properties makes use of the following notation:

$$\hat{\mu}(u) = P(T = P|u) \text{ with } \mu \equiv \int_U \hat{\mu}(u) dG(u) \tag{40}$$

Definition 5 *(i) Homogenous treatment effects within types (HTET):* $\hat{\Delta}_T(u) = \Delta_P$ *for all* u *for* $T \in \{P, N\}$. *(ii) Orthogonality of Verdicts and Treatment Effects with respect to other Unobservables (OVTU):* $\frac{\hat{\mu}(u)\hat{\Delta}_P(u)}{\mu} \neq \Delta_P \Longrightarrow \hat{\beta}^i(u) = \beta^i$; *and* $\frac{\hat{\mu}(u)\hat{\Delta}_N(u)}{\mu} \neq \Delta_N \Longrightarrow \hat{\alpha}^i(u) = \alpha^i$.

HTET is easy to explain: It suggests that the only factor that affects treatment effects is the subject's type. A slightly weaker version of this property can also be formulated by relaxing the requirment for inframarginal subjects, i.e., those who would never be treated by any judge and those who would be treated by every judge.

OVTU, on the other hand, is a more complicated concept. Intuitively, it requires unobservables, other than θ, that have an impact on treatment effects,

by type, to not have an impact on judges' (true or false) positive rates when they encounter that type, and vice versa. Thus, OVTU would be violated when there is a factor that is relevant to the judge's decision making, which is unobservable to the researcher but observable to the judge, and which affects the outcomes for some subjects differently when they receive a positive versus a negative verdict.

All three properties, HTET, OVTU, and DIU have similar implications, which I note as follows.

Remark 4 *(i)* $\Delta_Y^{i,j} = \mu(\beta^i - \beta^j)\Delta_P + (1-\mu)(\alpha^i - \alpha^j)\Delta_N$ *for all* i, j *if one of OVTU, HTET or DIU holds. (ii) ATM combined with one of OVTU, HTER or DIU implies that LATE estimates are valid.*

As noted in remark 4, all three properties imply the same relationship between the outcome differential observed between two judges, and the average treatment effects by type. Thus, when either of these conditions holds jointly with ATM, instrumental variables designs that exploit differences in judge propensities will return valid LATE estimates. However, when this is not true, absent other conditions that can act as replacements, one would need to rely on general monotonicity. Unfortunately, it is difficult to test general monotonicity, because it is caused by unobservables which do not generate easily identifiable implications. The same is not true for ATM and the three conditions in remark 4. These conditions have implications which can be tested if either miss rates or false hit rates are observable. The next section explains how one can test these assumptions, and how one can obtain both average treatment effects by type and valid LATEs, even when ATM is violated.

4.1 Tests of ATM and Condition 1

As discussed in Chan et al. (2022), when one has information regarding classifiers' miss rates, one can design a test that detects violations of ATM in cases where more traditional tests of monotonicity are unable to detect violations. This test relies on the observability of miss rates, which, in the current context corresponds to the false negative rate $\mu(1-\beta^i)$ for judge i. These, and similar rates that I refer to here, of course, are defined based on average rates as opposed to rates by unobservables (i.e. $\hat{\beta}^i(u)$), since the latter are unobservable to the researcher. Similar tests can also be designed when information regarding any of the four rates in (**??**) is observable, e.g., 'false hit rates' which equal $FP = (1-\mu)\alpha^i$, is observable. Thus, I define the difference in the positive rates in (**??**) for derivation purposes, as follows

$$\begin{aligned} \delta_T^{i,j} &\equiv TP^i - TP^j = \mu(\beta^i - \beta^j); \text{ and} \\ \delta_F^{i,j} &\equiv FP^i - FP^j = (1-\mu)(\alpha^i - \alpha^j) \end{aligned} \tag{41}$$

It is important to note that when either difference is observable, it follows that the other rate can also be treated as if it is observable. This is because the propensity difference for two judges is observable and is given by

$$\Delta_\Psi^{i,j} = \delta_T^{i,j} + \delta_F^{i,j} \tag{42}$$

Next, I explain how one can test both ATM and Condition 1 when either of these rates is observable, where the former test is proposed in Chan et al. (2022).

4.1.1 A Test of ATM

The test for identifying violations of ATM is proposed in Chan et al. (2022), and thus I simply reproduce by adopting the present notation.

Proposition 6 $ATM \implies \frac{\delta_F^{i,j}}{\delta_F^{i,j}+\delta_T^{i,j}} \in [0,1]$

Proof. Suppose $\delta_F^{i,j}, \delta_T^{i,j} \neq 0$. Then, $\frac{\delta_F^{i,j}}{\delta_F^{i,j}+\delta_T^{i,j}} > 0$, since, due to ATM, $sign(\delta_F^{i,j}) = sign(\delta_T^{i,j})$. Similarly, because $sign(\delta_F^{i,j}) = sign(\delta_T^{i,j})$, $\frac{\delta_F^{i,j}}{\delta_F^{i,j}+\delta_T^{i,j}} \leq 1$ iff $\frac{\delta_F^{i,j}+\delta_T^{i,j}}{\delta_F^{i,j}} = 1 + \frac{\delta_T^{i,j}}{\delta_F^{i,j}} \geq 1$, which holds. ■

4.1.2 A Test of Condition 1

Next, note that when DIU or OVTU holds, as noted in remark 4, the difference in outcomes can be expressed as:

$$\Delta_Y^{i,j} = \delta_T^{i,j}\Delta_P + \delta_F^{i,j}\Delta_N \tag{43}$$

Thus, for classifier pairs i, j, such that $\delta_M^{i,j}, \delta_F^{i,j} \neq 0$, one can specify the unobservable Δ_N as a function of the other unobservable Δ_P, denoted $\hat{\Delta}_N^{i,j}(\Delta_P)$ as

$$\bar{\Delta}_N^{i,j}(\Delta_P) = \frac{\Delta_Y^{i,j}}{\delta_F^{i,j}} - \frac{\delta_T^{i,j}}{\delta_F^{i,j}}\Delta_P \tag{44}$$

Here, $\bar{\Delta}_N^{i,j}(\Delta_P)$ expresses how large Δ_N would have to be for any given $\Delta_Y^{i,j}$, for the observed true and false positive rate differentials to produce the observed difference in outcomes associated with judges i and j. A similar relationship exists for all other pairs of judges for which the false and true positive rate differentials are not zero. Thus, if Condition 1 holds, then plots of (44) in Δ_N, Δ_P space for any pair of classifiers would have to intersect at the same Δ_N,Δ_P combinations, or, lie exactly on top of each other. The latter possibility complicates the notation necessary to describe a test that can be conducted to identify whether DIU holds, but can be accommodated by making simple adjustments. Thus, in order to specify the test in a more straightforward manner, I make the following assumption.

Assumption 1 *For any two non-identical pairs of judges i, j and k, m: (i) $|\delta_F^{k,m}| + |\delta_F^{i,j}|, |\delta_T^{k,m}| + |\delta_T^{i,j}| > 0$, and (ii) if $|\delta_F^{k,m}|, |\delta_F^{i,j}| > 0$ then $\frac{\delta_T^{k,m}}{\delta_F^{k,m}} \neq \frac{\delta_T^{i,j}}{\delta_F^{i,j}}$.*

With this assumption in place, a test for detecting violations of Condition 1, as well as a method for calculating Δ_N and Δ_P when Condition 1 holds, can be obtained as follows.

Proposition 7 *For any two non-identical pairs of judges, i, j and k, m, define*

$$T^{i,j;k,m} \equiv \begin{cases} \frac{\delta_F^{k,m}\Delta_Y^{ij} - \delta_F^{i,j}\Delta_Y^{k,m}}{\delta_F^{k,m}\delta_T^{i,j} - \delta_F^{i,j}\delta_T^{k,m}} & \text{if} \quad \delta_F^{i,j}, \delta_F^{k,m} \neq 0 \\ \frac{\Delta_Y^{ij}}{\delta_T^{i,j}} & \text{if} \quad \delta_F^{i,j} = 0 \\ \frac{\Delta_Y^{km}}{\delta_T^{k,m}} & \text{if} \quad \delta_F^{k,m} = 0 \end{cases} \tag{45}$$

If Condition 1 holds, then (i) there exists some T such that $T^{i,j;k,m} = T$ for all non-identical pairs i, j and k, m; and (ii) $\Delta_P = T$ and $\Delta_N = \bar{\Delta}_N^{i,j}(T)$ for any judge pair i, j.

Proof. The first expression in (45) is obtained immediately by using the equality in (44) and setting $\bar{\Delta}_N^{i,j}(\Delta_P) = \bar{\Delta}_N^{k,m}(\Delta_P)$. The remaining expressions are obtained by setting $\Delta_Y^{i,j} = \delta_T^{i,j}T$ and $\Delta_Y^{k,m} = \delta_T^{k,m}T$, respectively. ■

Proposition 7 specifies a very simple relationship that is implied by Condition 1. The key value in this proposition, namely $T^{i,j;k,m}$, is obtained by calculating the Δ_P that would be obtained by crossing $\bar{\Delta}_N(\Delta_P)$ curves for two different pairs of judges assuming Condition 1 holds. When Condition 1 holds, it follows that these curves would cross at the same Δ_P value, regardless of which judge pairs are chosen. Example 1, in section 4 below, illustrates this result graphically.

Part (ii) of proposition 7 notes that when Condition 1 holds, the average treatment effects by type (i.e., Δ_N and Δ_P) can be calculated as a function of observables. It is worth distinguishing these values from LATEs. While LATEs specify a weighted average of these treatment effects, Δ_N and Δ_P specify the true treatment effects by type. Thus, when Condition 1 holds, the true LATE will naturally lie in between the two average treatment effects by type. However, the specific weights attached to each treatment effect will depend on the judges positive rate differentials, as can be illustrated by the expression for the pairwise LATE produced, below.

Remark 5 *If DIU or OVTU holds, the true pairwise LATE associated with two classifiers i, j is given by*

$$L^{i,j} = \frac{|\delta_F^{i,j}|\Delta_N + |\delta_T^{i,j}|\Delta_P}{|\delta_F^{i,j}| + |\delta_T^{i,j}|} \tag{46}$$

Next, I provide examples illustrating how (44) and (45) appear when Condition 1 holds and when they are both violated.

4.2 Examples

To construct simple examples, I consider cases where $U = \{0, 1\}$ and denote their discrete probability distribution as $g(0) = g(1) = 0.5$, with $\mu(u) = 0.5$ for both u, $y_{v,T}(u) = 0$ for all t and u, $\Delta_N(0) = \Delta_N(1) = 0$, and $\hat{R}_N^i(0) = \hat{R}_N^i(1)$. I construct two examples within this set-up, one where Condition 1 holds but

both types of monotonicity are violated, and one where GM and Condition 1 are violated but ATM holds. The first example illustrates how one can obtain average treatment effects by type, even when estimated LATEs are biased. The second example illustrates how a monotonicity test based on miss rates may fail to detect violations of GM and thus fail to detect the invalidity of LATEs. In both examples I consider only four judges, and use tables to summarize each judge's propensities, in addition to the information necessary to express (44)

Example 3 *DIU and OVTU hold, but GM and ATM do not:* $\Delta_P = 0.5$ *with* $\hat{\Delta}_P(0) = 0$, $\hat{\Delta}_P(1) = 1$, $\mu(u) = 0.5$ *for* $u = 0, 1$, *and* $\hat{\beta}^i(0) = \hat{\beta}^i(1)$ *for all* i.

Table 1a: Measures of each classifier

	β^i	α^i	Y^i	Ψ^i
C_1	0.6	0.1	0.15	0.35
C_2	0.8	0.2	0.2	0.5
C_3	0.4	0.3	0.1	0.35
C_4	0.7	0.25	0.175	0.475

Table 1b: Pair Differentials and $\bar{\Delta}_N(\Delta_P)$

Pair	Δ_Y^{ij}	$\delta_T^{i,j}$	$\delta_F^{i,j}$	$\bar{\Delta}_N^{i,j}(\Delta_P)$
$(1,2)$	-0.05	-0.1	-0.05	$1 - 2\Delta_P$
$(1,3)$	0.05	0.1	-0.1	$-0.5 + \Delta_P$
$(1,4)$	-0.025	-0.05	-0.075	$\frac{1}{3} - \frac{2}{3}\Delta_P$
$(2,3)$	0.1	0.2	-0.05	$-2 + 4\Delta_P$
$(2,4)$	0.025	0.05	-0.025	$-1 + 2\Delta_P$
$(3,4)$	-0.075	-0.15	0.025	$-3 + 6\Delta_P$

First, note that the behavior of the four classifiers violates ATM (and therefore GM), as can be visually observed by plotting the α^i, β^i pairs of each judge as in figure 4a, below, since monotonicity would require all α^i, β^i pairs to lie to either the north-east or south-west of another pair. Plotting $\bar{\Delta}_N^{i,j}(\Delta_P)$ for the six pairings in table 1b in figure 4b, below, illustrates how these curves generate the same intersections in Δ_N, Δ_P space when Condition 1 holds.

Figure 4a

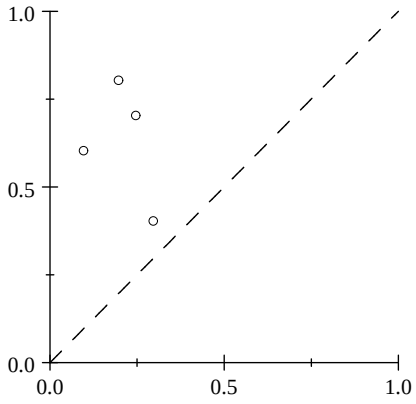
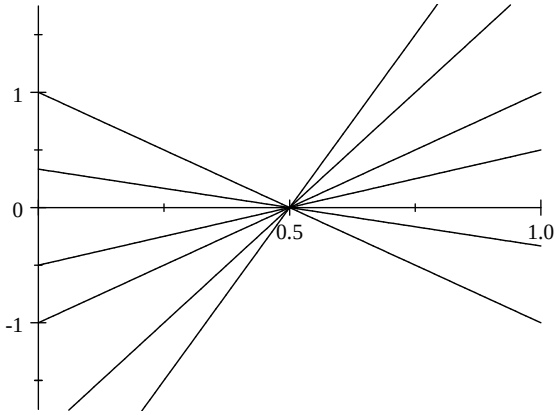


Figure 4b



The $\bar{\Delta}_N^{i,j}(\Delta_P)$ curves for the six pairwise matchings between the judges intersect at the average treatment effects by type, namely $\Delta_N = 0$ and $\Delta_P = 0.5$.

Example 4 *Condition 1 and GM are violated, but ATM holds:* $\Delta_P = 0.5$ *with* $\Delta_P(0) = 0$, $\Delta_P(1) = 1$, $\mu(u) = 0.5$ *for* $u = 0, 1$, *and* $\hat{\beta}^i(0) \neq \hat{\beta}^i(1)$ *for some* i.

Table 2a: Measures of each classifier

	$\hat{\beta}^i(0)$	$\hat{\beta}^i(1)$	β^i	α^i	Y^i	Ψ^i
C_1	0.6	0.2	0.4	0.1	0.05	0.25
C_2	0.55	0.45	0.5	0.2	0.1125	0.35
C_3	0.7	0.7	0.7	0.3	0.175	0.5
C_4	0.6	0.9	0.75	0.6	0.225	0.675

Table 2b: Pair Differentials and $\bar{\Delta}_N(\Delta_P)$

Pair	Δ_Y^{ij}	$\delta_T^{i,j}$	$\delta_F^{i,j}$	$\bar{\Delta}_N(\Delta_P)$
$(1,2)$	-0.0625	-0.05	-0.05	$\frac{5}{4} - \Delta_P$
$(1,3)$	-0.125	-0.15	-0.1	$\frac{5}{4} - \frac{3}{2}\Delta_P$
$(1,4)$	-0.175	-0.175	-0.25	$\frac{7}{10} - \frac{7}{10}\Delta_P$
$(2,3)$	-0.0625	-0.1	-0.05	$\frac{5}{4} - 2\Delta_P$
$(2,4)$	-0.1125	-0.125	-0.02	$\frac{9}{16} - \frac{5}{8}\Delta_P$
$(3,4)$	-0.05	-0.025	-0.15	$\frac{1}{3} - \frac{1}{6}\Delta_P$

Figure 5a plots the average positive verdict probability pairs (i.e., α^i, β^i pairs, with circles) and the $\alpha^i, \hat{\beta}^i(0)$ (marked with $\diamondsuit$ signs) and $\alpha^i, \hat{\beta}^i(1)$ pairs (marked with $+$ signs) to depict a graphical illustration of the fact that although ATM is satisfied, GM is violated. Because ATM is satisfied, it follows that $\bar{\Delta}_N^{i,j}(\Delta_P)$ curves are downward sloping for all i, j pairs. However, unlike in the case where Condition 1 holds, the curves do not intersect at a single point, as depicted in figure 5b, below.

Figure 5a

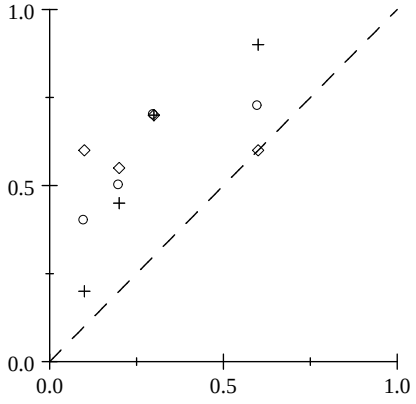
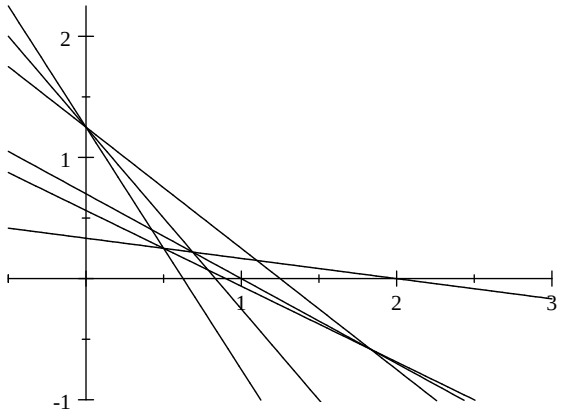


Figure 5b



Examples 3 and 4 together illustrate how global monotonicity may be violated in an undetectable manner, how violations of Condition 1 can be detected, and how average treatment effects by type can be calculated when Condition 1 holds.

5 Conclusion

A central condition in instrumental variables designs exploiting variation across judges' propensities is monotonicity. When judges differ in their abilities, there is no good reason to assume that this assumption will hold, and in fact, when

the signals that judges interpret to make decisions satisfy plausible properties, one would expect judge decisions to violate monotonicity. Moreover, even when estimated LATEs are valid, policy implications based on them may be misleading due to judges changing their behavior in response to changes in policies. These problems can be exacerbated when average treatment effects differ across the true type of subjects, which the judge attempts to identify correctly. These treatment effects can differ significantly from each other when subjects' learning and information updating processes differ across their true types. Inference problems caused by these complications can be circumvented by estimating the average treatment effect by subject type when factors unobservable to researchers but observable to the researcher satisfy certain properties described in this article. Whether these properties hold in important contexts is an empirical question. However, these conditions can be tested with less information than is necessary to test the types of monotonicity conditions that have been used as sufficient conditions to obtain valid LATEs.

References

[1] Angrist, J., Imbens, G., and D. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables" 91 Journal of the American Statistical Association 444-455.

[2] Bhuller, M. and H. Sigstad (2022) "Errors and Monotonicity in Judicial Decision-Making" 215 Eonomics Letters 110486.

[3] Chan, D., Gentzkow, M. and C. Yu (2022) "Selection with Variation in Diagnostic Skill: Evidence from Radiologists" 137 Quarterly Journal of Economics 729-783.

[4] Frandsen, B., Lefgren L., and E. Leslie, "Judging Judge Fixed Effects" NBER Working Paper no. 25528, 2019.

[5] Lundberg, A. and M. Mungan (2022) "The Effect of Evidentiary Rules on Conviction Rates" George Mason Law & Economics Research Paper No. 20-17.

[6] Miceli, T., Segerson, K., and D. Earnhart (2022) "The Role of Experience in Deterring Crime: A Theory of Specific versus General Deterrence" Economic Inquiry 1-21 Available from: https://doi.org/10.1111/ecin.13083

[7] Mungan, M. (2017a) "Reducing Crime through Expungements" 137 Journal of Economic Behavior an Organization 398-409.

[8] Mungan, M. (2017b) "The Certainty versus the Severity of Punishment, Repeat offenders, and Stigmatization" 150 Economics Letters 126-129.