Post Hoc Interventions in Criminal Sentencing:
An Empirical Thought Experiment

Erik J. Girvan

In Gunnemyr, Mattias & Jönsson, Martin L. (2023) *Post Hoc Interventions: Prospects and Problems.* Lund: Department of Philosophy, Lund University. https://doi.org/10.37852/oblu.184

ISBN: 978-91-89415-60-7 (print)
978-91-89415-61-4 (digital – pdf)
978-91-89415-62-1 (digital – html)

DOI: https://doi.org/10.37852/oblu.184.c504

Published by the Department of Philosophy, Lund University.
Edited by: Mattias Gunnemyr and Martin L. Jönsson
Cover layout by Cecilia von Arnold, Pufendorf Institute for Advanced Studies

4

Post Hoc Interventions in Criminal Sentencing

An Empirical Thought Experiment

Erik J. Girvan[1]

Abstract. Post Hoc Interventions (PHIs) are approaches for reducing the impact of discriminatory ratings, evaluations, or other decisions by correcting statistically for the impermissible discrimination before applying the decisions. Scholars have proposed a set of empirical criteria that are theoretically necessary for implementation of PHIs. In this paper, I conduct an empirical thought experiment to examine how the criteria, along with a normative consideration derived from U.S. anti-discrimination law, relate to conditions in an actual case: Application of PHIs to adjust for potential ethnic biases in criminal sentencing outcomes. Results suggest that, while the criteria may not all be present in their strong form, allowing for reasonable inferences, in many circumstances they can be likely satisfied in practice.

Introduction

A core tenant of the rule of law is that legal decisions ought not to be decided arbitrarily. Rather, following the Aristotelian notion of justice, like cases should be decided alike and different ones differently. Deciding which attributes of cases determine whether they are like or different is a normative question. Assessing whether the attributes are present in the circumstance of a particular case is an empirical one.

In the United States and elsewhere, there is an anti-discrimination norm embodied in legal doctrine (Girvan, 2020; Liebman, Butler, Buksunski, 2021). The norm provides that one class of attributes that ought not to be used to determine if cases are like is the race, ethnicity, or sex of those involved, along

[1] Erik J. Girvan, Associate Professor at the University of Oregon School of Law, University of Oregon.

with other protected attributes. Individuals who can show that they were treated differently by government officials, employers, or businesses based on one of these attributes can thus obtain equitable or financial relief.

Decision-makers who share the anti-discrimination norm or who wish not to be legally liable for violating it adopt a range of strategies to avoid making decisions based on the protected attributes (see e.g., Hassen et al., 2021; Madva, 2020). Most commonly these efforts are preventative, targeting factors (e.g., explicit and implicit bias) thought to contribute to impermissible, discriminatory decision-making. However, the preventative efforts have a mixed record of success (Lai et al, 2014; Lai et al, 2016; Mcintosh, Smolkowski, Gion, et al, 2020). Sometimes the efforts reduce disparities related to the protected attributes. Often, they do nothing. Occasionally they produce backlash effects, making the disparities worse (for a review see Tellhed, this volume).

In addition or as an alternative to preventative approaches, Jönsson and colleagues (Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017) suggest that harm from discriminatory decision-making may be mitigated after the fact using statistical methods to directly correct decisions for the extent to which protected attributes like race, ethnicity, or sex influenced their outcome, an approach they refer to as *post hoc interventions* (PHIs). In addition, they identify a set of empirical conditions thought to be necessary for use of PHIs.

The goal of this paper is to conduct an empirically grounded thought experiment into the viability of PHIs in practice. In particular, building on the findings reported in Girvan and Marek (2023), I use PHIs to correct for racial and ethnic disparities in the extent to which White and Hispanic individuals are sentenced to prison, as compared to jail or probation, for violations of criminal laws. In doing so, I apply the empirical requirements for PHIs discussed by Jönsson and Bergman (2022), along with an additional normative limitation on steps one may take to correct for disparities based on protected attributes, and discuss the implication for PHIs and flexibility in the specified conditions in practice.

Conditions for Post Hoc Interventions

PHIs are adjustments to ratings, evaluations, or other assessments, r, of a latent characteristic, c, that has been identified as a legitimate basis for decision-making. Their use involves three basic steps. First, prior ratings, r_0, are examined to determine whether they differ impermissibly based on protected attributes of the individuals being evaluated. If individuals with certain of the attributes, e.g.,

Men, have been assigned higher evaluations on r_0 than those with comparison attributes, e.g., Women, under conditions in which members of the two groups can be assumed to have the same distribution of the latent characteristic c_0, then the group difference in r_0 is assumed to reflect impermissible use of the attribute. Second, a precise, incremental, quantitative correction, v, is statistically identified that, when applied to r_0, produces the same evaluations for individuals irrespective of the attribute. Third, v is applied to future evaluations ($r_{1\text{-}n}$) and the result, $r_{1\text{-}n} + v$, used to determine the decision outcome.

Jönsson and colleagues (Jönsson & Bergman, 2022; Jönsson & Sjödahl, 2017, see also Jönsson, this volume) discuss the empirical conditions that must, in theory, be present in order to use PHIs. They are, restated and summarized in my terms:

1. *Interval scale ratings*. To be able to calculate and apply correction, v, to ratings, r_0, r_0 must be on an interval scale.

2. *Low error*. To be able to justify application of correction, v, underlying estimates of differences in r_0 must be based on a large enough sample of r_0 such that the extent of error in estimates of group differences is sufficiently narrow.

3. *No unknown differences*. To be able to justify the inference that differences in r_0 are attributable to impermissible consideration of a protected attribute, there must either no or known differences in c_0 based on that attribute.

4. *Constant bias over time*. To be able to justify the inference that application of v to $r_{1\text{-}n}$ is corrective of impermissible consideration of a protected attribute in those future evaluations, the magnitude of the differences in r_0 and $r_{1\text{-}n}$ based on the attribute must not vary systematically.

5. *Same categorization as bias*. To be able to correct for impermissible consideration of a protected attribute using v, individuals being rated must be categorized in the same way on the attribute in the PHI process as they were by the evaluators who produced r_0.

6. *Same contingencies as bias*. To be able to correct for impermissible consideration of a protected attribute using v, the PHI process must incorporate any contingencies regarding differences in r_0 based on the attribute (e.g., intersectionality between two or more protected attributes, interactions between a protected and permissible attributes).

7. *Same relationship as bias*. To be able to correct for impermissible consideration of a protected attribute using v, v must reflect the relationship (e.g., linear, non-monotonic) between the attribute and r_0 in the evaluations.

In addition to the empirical limitations, there are numerous potential normative considerations regarding the conditions under which one ought or ought not to directly correct for disparities related to protected characteristics in evaluations, ratings, or other assessments. Here I consider one.

1. *Cure not worse than the disease*. The anti-discrimination norm provides that decisions about people ought not be directly impacted by their status with respect to their race, ethnicity, sex, or other protected attributes. By adjusting r_0 based on such characteristics, PHIs are arguably doing just that. Under the norms embodied in U.S. anti-discrimination law, the adjustments are justified as a corrective measure only to the extent that we are sure that the group difference was caused by impermissible consideration of the attributes, e.g., racism or sexism of decision-makers (Girvan, 2020). They may not, however, be justified if the differences are attributable to random error in the sample or extrinsic factors that are causally related to c_0 and also happen to be correlated with the protected attributes (Chemerinsky, 2014; Rutherglen, 2009). To the extent that r_0 differs based on protected attributes of the individuals being evaluated for a reason other than impermissible consideration of the attributes, deliberately applying v to $r_{1\text{-}n}$ based on an individual's status with respect to protected attributes may thus be regarded as itself a violation of the anti-discrimination norm.

Application of PHIs to Criminal Sentencing Decisions

Could PHIs be used to correct for racial disparities in criminal sentencing decisions? As an empirical thought experiment, I apply the PHI approach to adjust for ethnic disparities in a set of actual sentencing decisions. In doing so, I compare and contrast the conditions of the cases of criminal sentencing to the empirical conditions identified as necessary for PHIs as well as the normative consideration and identify implications of any similarities or differences.

Sample of Criminal Sentencing Decisions

For the empirical thought experiment, I used a sample of records of sentencing decisions regarding 222,035 unique sentenced offenses (USOs)[2] committed by

[2] USOs are the most serious concurrently sentenced offenses for each individual. An individual who was simultaneously sentenced for four offences the sentences for each of which were to be served concurrently would have only one – the most serious sentenced offense – in the sample as one USO. If an individual completed a sentence and then committed and were sentenced for

195,854 people who were ultimately incarcerated in the State of Oregon at any point between 2004 and 2018 and belonged to one of three racial/ethnic categories. White-White individuals (N=162,742; USOs=184,976) were those identified by actors in the legal system as White (non-Hispanic) and predicted, using validated estimates, to self-identify as White. Hispanic-Hispanic individuals (N=21,101; USOs=23,983) were identified by actors in the legal system as Hispanic (any race) and predicted, using validated estimates, to self-identify as Hispanic. White-Hispanic individuals (N=12,011; USOs=13,076) were identified by actors in the legal system as White and predicted, using validated estimates, to self-identify as Hispanic.

In the United States, sentences for more severe crimes are generally served in state-run prisons (i.e., longer-term, more secure facilities). By comparison, sentences for less serious offenses are generally to jail (i.e., short-term, locally operated facilities), probation, or a combination of the two. Consistent with the distinction, in the sample, 60,240 USOs resulted in sentences to prison and 161,795 sentences to jail, probation, or both.

In Girvan and Marek (2023), my collaborator and I analyzed this sample of criminal sentencing decisions to determine whether race and ethnicity of the individuals sentenced impacted the likelihood of their being sentenced to prison as compared to jail/probation. To summarize, psychological theory indicates that, for group-based biases to impact decisions, decision-makers must first identify and categorize target individuals as members of the relevant group. Accordingly, we reasoned that, to the extent group-based biases impacted sentencing decisions, there would only be sentencing differences based on perceived race/ethnicity, not self-identified race/ethnicity where the two differed: After accounting for legally relevant factors, individuals perceived by those in the criminal justice system as Hispanic would be more likely to be sentenced to prison than similarly situated individuals perceived to be White. However, sentences of individuals misperceived as White but who self-identified as Hispanic would not differ from those of individuals accurately perceived as White. Our findings were consistent with the predictions. Even after controlling for crime severity and criminal history, individuals who were accurately labeled as Hispanic in criminal justice records (Hispanic-Hispanic) were nearly twice as likely to be sentenced to prison as those who were accurately labeled as White [White-White; Odds Ratio: 1.95 (95% CI: 1.86, 2.04)]. By comparison, individuals who were mis-perceived in

---

another offence, or if they committed two offenses the sentences for which were served consecutively, then they would appear in the dataset twice, once for each USO.

criminal justice records as White but who, based on validated estimates, self-identified as Hispanic (Hispanic-White) had the same likelihood of prison sentences as those who were accurately perceived to be White [Odds Ratio: 1.01 (95% CI: 0.94, 1.07)].

The empirical thought experiment takes a step further from the findings in Girvan and Marek (2023) by asking: What would it look like to use PHIs to attempt to correct for the observed disparity? However, the PHI approach uses past estimates of disparities in ratings r_0 to create a corrective function, v, to be applied to future ratings, $r_{1\text{-}n}$. Accordingly, rather than using all of the data for r_0 and $r_{1\text{-}n}$, for the thought experiment, I split the sample into sentences of USOs up to and including 2014 (N=168,290), which I treated as r_0, and those after 2014 (N=53,745), which served as $r_{1\text{-}n}$.

PHI Steps

PHI Step 1: Identification of Disparities in r_0

The first step in PHIs involves use of extant data regarding ratings, r_0, of a latent characteristic, c_0, that has been identified as a legitimate basis for decision-making to determine whether they differ impermissibly based on protected attributes of the individuals being evaluated. Here, as is typical in the U.S., criminal sentencing decisions in Oregon are made with reference to a set of sentencing guidelines designed to assign longer and more punitive sentences to what I will refer to as more reprehensible behavior, c_0. The guidelines operationalize reprehensibility and provide for the duration of criminal sentences using two underlying considerations: The severity of the offence committed and the extent of the criminal history of the individual being sentenced (Or. Admin. R. 213-004-0001). At the intersection of any level of offense severity and criminal history, the guidelines provide a presumptive sentencing range within which the sentencing judge has discretion to choose the appropriate sentence (Or. Admin. R. 213-004-0001; Or. Admin. R. 213-005-0007). The sentencing judge may depart from a presumptive sentence range, but only upon a finding of "substantial and compelling reasons" to do so (Or. Rev. Stat. § 137.671; Or. Admin. R. 213-008-0001). Such departures may be dispositional (imposing probation when the presumptive sentence is prison or vice versa) or durational (diverging from the presumptive sentence as to the term; Or. Admin. R. 213-003-0001(6), (8)). Thus, in theory, adhering to sentencing guidelines, individuals with comparable criminal histories who commit similarly severe offenses should receive like sentences, r_0 (Mitchell, 2017).

Table 1: Logistic Regression Coefficients and Odds Ratios Indicating Likelihood of Sentences to Prison Compared to Jail and/or Probation by Offender Race and Ethnicity.

	Coefficients		**Odds Ratios**	
Intercept	-3.356	[-3.536, -3.176]	.04	[0.03, 0.04]
Race/Eth. (White-White)				
White-Hispanic	-.026	[-.108, .056]	.97	[0.90, 1.06]
Hispanic-Hispanic	.724	[.672, .776]	2.06	[1.96, 2.17]
Psudo-R2		.956		

Note. Cell values are logistic regression coefficients (first column) or corresponding odds ratios (third column) followed, in brackets, by the 95% confidence intervals. All p-values are less than .001 except that for White-Hispanic (p = .530). Coefficients and odds ratios for legally relevant factors omitted from table.

To assess whether there was an impermissible difference in sentences of USOs, r_0, based on the perceived race and ethnicity of the individuals being sentenced, I fit a logistic regression model that includes the legally relevant factors that should, under the law, determine the type of sentence: Indicators of the individuals' offense history and offense severity. To this I added the sentenced individuals' sex and race/ethnicity, described above (see Girvan & Marek, 2023). To account for potential impacts of lack of independence, p-values and confidence intervals for coefficients were calculated using cluster-robust standard errors.

The relevant portion of the results of the analysis are given in Table 1. Effectively replicating the results of Girvan and Marek (2023), they indicate that sentencing decisions made from 2004 to 2014 regarding USOs of individuals who were perceived to be Hispanic (i.e., Hispanic-Hispanic) were approximately twice as likely to result in a sentence to prison than decisions regarding USOs by legally similarly situated individuals accurately perceived to be White (i.e., White-White). By comparison, decisions about USOs committed by individuals perceived to be White but who, based on validated estimates, would self-identify as Hispanic (i.e., White-Hispanic) did not differ from those of White-White individuals.

PHI Step 2: Calculate Correction v

The second step of the PHI-process is to calculate a precise, incremental, quantitative correction, v, that, when applied to r_0, produces the same ratings

for individuals irrespective of their status on the protected attribute. Here we can use the value of the coefficient for Hispanic-Hispanic individuals in the model from step 1: .724.

PHI Step 3: Apply Correction v to $r_{1\text{-}n}$

The third step is to apply v to future evaluations, $r_{1\text{-}n}$ and use the result, $r_{1\text{-}n} + v$, to determine the decision outcome. To do so, I used the coefficients from the logistical model in Step 1 to generate predicted log-odds of a sentence to prison for each USO decided after 2014, i.e., the as $r_{1\text{-}n}$ dataset. I then subtracted the adjustment v of .724 from the log-odds of predicted sentences for USOs by Hispanic-Hispanic individuals. The adjusted log-odds were then used to predict sentencing decisions for all USOs, with those having a log-odds greater than 0, the equivalent to an odds greater than 1, being to prison.

To illustrate the impact of the adjustment, Table 2 provides the actual sentences (top two rows), sentences that would be predicted by the unadjusted r_0 model coefficients (middle two rows), and sentences predicted by the adjusted coefficients (bottom two rows). Notably, comparison of the actual sentences to the un-adjusted predicted sentences shows that the only group for which the un-adjusted coefficients over-predict prison sentences is Hispanic-Hispanic offenders. Application of the adjustment corrects this, bringing the predicted sentences more in line with those for the other groups. Comparing the predictions of the unadjusted and adjusted models thus suggests that application of the adjustment results, depending on the baseline, in

Table 2: Outcomes of Actual, Unadjusted Predicted, and Adjusted Predicted Sentences of USOs Made after 2014

		White-White	**Hispanic-Hispanic**	**White-Hispanic**
Actual Sentences	Prison	11,457 (.213)	1,947 (.036)	927 (.017)
	Jail/Probation	33,324 (.620)	3,422 (.064)	2,668 (.050)
Predicted Sentences	Prison	11,116 (.207)	2,192 (.041)	804 (.015)
	Jail/Probation	33,665 (.626)	3,177 (.059)	2,791 (.052)
PHI Adjusted Predicted Sentences	Prison	11,116 (.207)	1,695 (.032)	804 (.015)
	Jail/Probation	33,665 (.626)	3,674 (.068)	2,791 (.052)

Note. Cell values are counts followed by proportions of all sentences.

approximately 250 to 500 fewer prison sentences for USOs by Hispanic-Hispanic individuals sentenced post-2014, equivalent to about 5 to 10% of the USOs sentenced for this group.

Comparison of Example PHI to Identified Empirical and Normative Requirements

Interval Scale Ratings

Jönsson and Bergman (2022) specify that PHIs must be applied to interval ratings in order to be able to calculate a corrective function. As with many threshold decisions, the latent characteristic *c* upon which sentencing decisions are based, level of reprehensibility, may be thought of as a continuous, interval- (or even ratio-) level construct. Even so, when conceptualized as ratings, *r*, decisions regarding the nature of a criminal sentence are ordinal, representing a dichotomous decision to sentence an individual to prison, if sufficiently reprehensible, or jail/probation, if not. Consistent with the logistic regression approach used in the example, when such dichotomous decisions are aggregated, calculating a corrective function for them based on the log-odds of prison compared to jail or probation, which is on an interval scale, is straightforward. In practice, however, application of the corrective function to individual future decisions, $r_{1\text{-}n}$, is challenging. Judges do not issue their sentencing decisions in log-odds of prison and thus their decisions cannot be directly corrected in this way.

One alternative approach, used in the example, is to use the coefficients from the model fit on prior sentencing decisions, adjusted with *v*, to predict types of sentences for individual cases as they arise. This approach differs from the prototypical PHI, however, in that adjustments are not made directly to judges' ratings in the new cases. Indeed, if the predicted sentencing decisions are viewed as "correct," then, once the coefficients and adjustments are calculated, the decision process can be automated and judicial ratings in new cases are not actually required at all. To the extent that this is viewed as methodologically or normatively problematic, one could use a hybrid system in which judges continue to make sentencing decisions in parallel with adjusted predicted decisions generated from all prior sentencing decisions. Where the two differ, the predicted decision will be used, the judge notified that a protected attribute may have influenced the decision and invited to reconsider, or another layer of processes added such as supplemental review by a panel. In any of these scenarios, judicial decisions would govern in most cases and, where they did

not, at a minimum, they would continue to influence sentences indirectly through inclusion in the sample used to generate coefficients in future estimation (Step 1) or adjusted prediction (Step 3) models.

Low Error

A second requirement for PHIs is that the underlying estimates of differences in r_0 must be based on a large enough sample such that the extent of error in estimates of group differences is sufficiently narrow to be usable. In the example, there are ample prior sentencing decisions to make reliable estimates of the influence of protected attributes on decisions. Indeed, the range of the 95% confidence interval around the coefficient estimate for the influence of perceived Hispanic ethnicity is equivalent to just .02 of the standard deviation in log-odds.

No Unknown Differences

The third requirement is that, for the inference that observed differences in r_0 are attributable to impermissible consideration of a protected attribute to be valid, there must be either no or known differences in c_0 based on that relevant protected attribute. In practice, this condition will not be met, except possibly in circumstances in which no judgment is required to operationalize the latent constructs that form the bases of the ratings being examined and no discretion afforded to raters interpreting or making ratings based on measures of them. In practice, however, there is also error and uncertainty in the measurement of nearly all latent characteristics and discretion in processes used to generate ratings from them. Accordingly, the requirement should turn on either (a) an assessment of whether the judgment conditions are such that an attribution of impermissible use of a protected attribute is a reasonable inference regarding the observed group difference or (b) application of a norm of presuming no unknown differences between groups, absent sufficient evidence to the contrary.

With respect to the sentencing decision example, the weaker requirement of a reasonable inference is satisfied in two ways. The first and perhaps most generalizable of the ways is that the institution on whose behalf the judges are making the decisions, the Oregon criminal justice system, had the opportunity to and did specify in advance the factors that ought to determine the outcome of the sentencing decisions: The severity of the USOs or the criminal record of the offender. Moreover, the influence of these factors was accounted for in the first step of the PHI, which showed that approximately 95% of the variance in

prison sentences was explained by them. Accordingly, it is a reasonable inference that observed differences between groups based on their status on a protected attribute stems from impermissible consideration of the attribute or an associated characteristic that ought not to impact the decision. Second and perhaps not as generalizable, the difference in sentencing decisions in the example is consistent with the predictions of psychological theory regarding the conditions under which group-based stereotypes and attitudes tend to influence decisions. And, while the correlational nature of the analysis precludes a strong inference of causality, any alternative explanation for the sentencing differences would also have to consider the fact that they are associated with perceived, but not self-identified, ethnicity.

Constant Bias over Time

Fourth, for a corrective function based on past ratings to accurately adjust for the impacts of impermissible consideration of protected attributes in future ratings, the magnitude of the impact must be relatively consistent over time. As with the requirement of no or known group differences, in practice it will often be impossible to know exactly the extent to which impermissible consideration of protected attributes changed over time. Given sufficient longitudinal data, however, it is relatively easy to determine whether differences in decisions associated with protected attributes remain relatively consistent, supporting a reasonable inference that the impact of potential impermissible influence of consideration of them on the decisions is also consistent.

To illustrate, I separately re-ran the logistic regression model on sentencing decisions made during four different time frames: 2004 to 2006, 2007 to 2009, 2010 to 2012, and 2013 to 2014. Results, given in Table 3 (next page), suggest that the magnitude of the increased likelihood of USOs by Hispanic-Hispanic individuals being sentenced to prison as compared to those by legally similarly situated White-White individuals rose over the first three periods and then decreased. Moreover, the change is sufficiently large that the highest of the coefficients and associated adjustments, .881, falls outside of the 95% confidence intervals for the other time periods.

To illustrate the impact of the change, we can repeat our PHI process three times, treating the earlier time period, e.g., 2004 to 2006, as r_0 from which we computer the adjustment v and the subsequent one, e.g., 2007 to 2009, as the $r_{1\text{-}n}$ to which we apply it. The result would be that, in the first two times we used PHIs, the adjustment would under compensate for the influence of

Table 3: Logistic Regression Coefficients and Odds Ratios Indicating Likelihood of Sentences to Prison Compared to Jail and/or Probation for USOs by Hispanic-Hispanic Individuals by Groups of Years Sentenced.

	Hispanic-Hispanic Coefficients		**Hispanic-Hispanic Odds Ratios**	
2004 - 2006	.634	[.532, .737]	1.89	[1.70, 2.09]
2007 - 2009	.724	[.631, .817]	2.06	[1.88, 2.26]
2010 - 2012	.881	[.779, .984]	2.41	[2.18, 2.68]
2013 - 2014	.672	[.540, .803]	1.96	[1.72, 2.23]

Note. Cell values are logistic regression coefficients (first column) or corresponding odds ratios (third column) followed, in brackets, by the 95% confidence intervals. All p-values are less than .001. Other coefficients and odds ratios omitted from table.

protected attributes on ratings in the subsequent period. By comparison, the third time the adjustment would over-compensate. How much such under- or over-compensation is acceptable may be context specific, depending on empirical considerations related to factors like the overall stability in the ratings themselves and normative considerations regarding the implications of incremental changes in the ratings. For example, where ratings fluctuate considerably over time but where incremental changes to ratings have a low impact on outcomes of others, e.g., where the decisions outcomes are independent as when assigning grades based on absolute performance, then instability may be less of a concern. In the context of the thought experiment, because sentencing decisions are relatively independent, i.e., adjusting the decision so that someone who would have gone to prison instead goes to jail or serves probation does not require that someone else who would have gone to jail or served probation to now serve a prison sentence, the level of instability observed here may be acceptable.

Same Categorization as Bias, Same Contingencies as Bias, and Same Relationship to Bias

The fifth, six, and seventh requirements for PHIs I identify each capture a type of complexity in the ways in which, as a result of social psychological processes, raters' impermissible consideration of protected attributes may impact r_0: Raters may categorize individuals based on protected attributes differently than would others or the individuals themselves, raters' decisions

may be influenced by the interactions between several protected attributes or protected attributes and characteristics of the rating situation, and the influence of protected attributes on r_0 may otherwise be non-linear. The more accurately the complexities are modeled in the PHI process, the more accurately v will be able to correct for bias when applied to $r_{1\text{-}n}$. As with the other requirements, if interpreted strictly, in practice, given variation in human perceptions, differences in the subjective salience of particular socially defined attributes, and the conditional nature of some biases, it will rarely be completely satisfied in circumstances involving room for interpretation or discretion. However, if viewed as requirements that PHIs may be done in circumstances where it is reasonable to infer that the primary influence of the protected attributes on ratings can be sufficiently similarly captured in the PHI process, then it may only limit some applications of PHIs in which there is particular reason for concern. To illustrate, for each requirement, I consider the example PHI in criminal sentencing in light of some potential ways in which estimation and application of v may differ from the original impacts of impermissible consideration of protected attributes on r_0.

First, for raters' attitudes and stereotypes regarding protected attributes to impact their decisions, the raters must identify someone based on their status in relation to the attributes (Rees, Ma, & Sherman, 2020). Where someone's status as to a protected attribute is difficult to observe reliably, data sources based on self-reported status regarding the attribute may differ from those based on perceived status. For example, in the U.S., research results indicate that individuals who identify as Hispanic or Native American tend to be mistaken for White (Girvan & Marek, 2023). In the Oregon Department of Corrections database, race and ethnicity of offenders were based on the perceptions of race and ethnicity by officials in the criminal justice system, such as the arresting law enforcement officers. In such circumstances, using self-identified race and ethnicity in PHIs, by, for example, asking inmates, job applicants, or others subject to decision-making to identify their own race and ethnicity, rather than recording the attributes perceived by the raters themselves, may result in inaccurate adjustments.

To illustrate, I re-ran the logistic regression model for Step 1 of the PHI using only the validated estimates of self-identified race and ethnicity rather than perceived values. The result indicated less of a difference between sentences of USOs by Hispanic and White individuals [$\beta = .494$ (95% CI: .524, .625); Odds Ratio: 1.78 (95% CI: 1.69, 1.87)], and thus that less of an adjustment would be needed than with the model using perceived race and ethnicity. If the coefficient and self-categorization approach were used to make

Table 4: Logistic Regression Coefficients and Odds Ratios Indicating Likelihood of Sentences to Prison Compared to Jail and/or Probation by Offender Race and Ethnicity, Sex, and the Interaction Between Them.

	Coefficients		Odds ratios	
Intercept	-3.363	[-3.543, -3.183]	.04	[0.03, 0.04]
Race/Eth. (White-White)				
White-Hispanic	.345	[.167, .524]	1.41	[1.18, 1.69]
Hispanic-Hispanic	.364	[.167, .561]	1.44	[1.18, 1.75]
Sex (Female)				
Male	.553	[.499, .607]	1.74	[1.65, 1.84]
White-Hispanic x Male	-.470	[-.670, -.271]	.63	[.51, .76]
Hispanic-Hispanic x Male	.384	[.180, .588]	1.47	[1.20, 1.80]
Psudo-R2		.956		

Note. Cell values are logistic regression coefficients (first column) or corresponding odds ratios (third column) followed, in brackets, by the 95% confidence intervals. All p-values are less than .001. Coefficients and odds ratios for legally relevant factors omitted from table.

adjustments in step 3, the result would be that the PHI would under correct $r_{1\text{-}n}$ for self-identified individuals who were perceived to be Hispanic and overcorrect those who were not.

The same potential problem may be extended further to protected attributes that are treated as categorical but the identification and influence of which varies continuously. One example of this is Afrocentric features, i.e., the extent to which people appear closer to stereotypes of the phenotype of individuals of African descent. Research on racial bias in sentencing in the U.S. indicates that people who have more Afrocentric features tend to receive harsher sentences than those with less Afrocentric features (Burch, 2015; King & Johnson, 2016). Under some circumstances, other potentially protected attributes like age may also influence judgments primarily continuously, thus making it important to capture raters' subjective perceptions of the characteristic directly.

Turning to contingencies, results of a substantial body of research suggests that the operation of stereotypes and attitudes regarding people based on their

Table 5: Outcomes of Actual, Unadjusted Predicted, and Adjusted Predicted Sentences of USOs Made after 2014

		White-White	Hispanic-Hispanic	White-Hispanic
Actual Sentences	Prison	11,457 (0)	1,947 (0)	927 (0)
	Jail/Probation	33,324 (0)	3,422 (0)	2,668 (0)
Predicted Sentences	Prison	11,114 (-2)	2,174 (-18)	781 (-23)
	Jail/Probation	33,667 (+2)	3,195 (+18)	2,814 (+23)
PHI Adjusted Predicted Sentences	Prison	8,262 (-2,854)	1,473 (-222)	700 (-104)
	Jail/Probation	36,519 (+2,854)	3,896 (+222)	2,895 (+104)

Note. Cell values are counts followed by change from original PHI values in Table 2.

attributes often interact or intersect with one another (Crenshaw, 2017; McCall, 2005). Stereotypes of or attitudes towards men and women, for example, may be qualitatively different than those for White men, Black men, White women, and Black women. Depending on the circumstances, understanding which cluster of attributes were salient to raters can be difficult because of the complexity of the interactions.

To illustrate the potential effects of intersecting attributes, I re-ran the logistic regression model for Step 1 of the PHI, adding the interaction terms between race and ethnicity and sex. The relevant results are given in Table 4. They indicate that the race and ethnicity differences from the original model are largely driven by the likelihood of USOs by Hispanic-Hispanic men resulting in a sentence to prison rather than jail or probation, which is higher than that for USOs by individuals of any other combination of race and ethnicity or sex.

To assess how much difference the outcomes of PHIs using a model that adjusts for significant effects of race and ethnicity, sex, and their interaction terms rather than just race and ethnicity, I used the model with interaction terms to predict un-adjusted and adjusted sentences. The results, in Table 5, provide the number of actual, predicted, and adjusted predicted sentences along with the change from these values in the original PHI (see Table 2). Review of the table shows that addition of the interaction of race and ethnicity and sex to the model did not dramatically change the unadjusted predictions of the model

(middle two rows). However, application of the adjustments resulted in a substantial reduction in the overall number of prison sentences for each group. In addition to interactions between attributes, social psychological theory also indicates that the influence of stereotypes and attitudes also tend to be moderated by features of a decision situation. For example, attitudes or stereotypes associated with protected attributes are more likely to affect decisions that are discretionary or in which the "correct" outcome is unclear, such as when there is some ambiguity or uncertainty in the decision criterion or an exercise of judgement required in order to make a decision (Girvan, 2016; see also Bushway & Forst, 2013; Bushway & Piehl, 2001). Sentencing guidelines were enacted, in part, to reduce racial disparities by limiting judicial discretion (Stith & Koh, 1993). Even with sentencing guidelines, however, judges retain some discretion on the margins to depart from guidelines and can impose sentences of different severity or divert individuals into alternatives like probation and rehabilitative programs. Judges deciding to depart from the guidelines generally do so based on some consideration of subjective factors such as rehabilitative potential or ties with the community (Painter-Davis & Ulmer, 2020).

To illustrate with the example sentencing decisions, Table 6 gives the distribution of the raw number of sentencing outcomes and that would be adjusted by the PHI process (see Table 2) in terms of the sentencing guidelines. Consistent with psychological theory, the adjustments are not randomly distributed across the sentencing grid but rather tend to be concentrated in areas of the guidelines near the threshold for prison or jail and probation sentences. For example, the largest proportion of adjustments (26% of the total) occurred for USOs classified as fairly severe, i.e., 8 out of 11 on the Crime Seriousness Scale, with 11 being the most serious, committed by Hispanic-Hispanic individuals who were at the lowest two lowest levels of the Criminal History Scale, i.e., individuals with no record of serious crime as an adult. The second largest proportion of adjustments (16%) were for USOs classified as moderately severe (6 on the Crime Seriousness Scale) committed in individuals who had at least one prior felony involving harm to a person. By comparison, the PHI adjustments did not change the outcomes of any sentences for USOs involving the most serious crimes, i.e., those at 10 or 11 of the Crime Seriousness Scale, for which prison sentences are effectively the "correct" outcome. Similarly, the PHI adjustments resulted in only a small number of changes to USOs for crimes very low on the Crime Seriousness Scale, those for which the "correct" outcome is a combination of jail and probation.

Table 6: Number and Proportion of Total Sentencing Decisions Regarding Hispanic-Hispanic Offenders that Differ Between Actual and PHI-Adjusted Outcomes by Location on the Sentencing-Guidelines Grid.

		Criminal History Scale									
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**X**
Crime Seriousness Scale	**11**	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	10	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	9	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	10 (.02)	16 (.03)	0 (0)
	8	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	5 (.01)	19 (.04)	32 (.06)	107 (.20)	0 (0)
	7	0 (0)	8 (.01)	5 (.01)	18 (.03)	3 (.01)	6 (.01)	5 (.01)	8 (.01)	17 (.03)	0 (0)
	6	14 (.03)	14 (.03)	26 (.05)	57 (.11)	3 (.01)	5 (.01)	3 (.01)	14 (.03)	26 (.05)	0 (0)
	5	1 (0)	1 (0)	2 (0)	1 (0)	4 (.01)	4 (.01)	3 (.01)	1 (0)	1 (0)	0 (0)
	4	6 (.01)	6 (.01)	2 (0)	2 (0)	1 (0)	0 (0)	2 (0)	0 (0)	1 (0)	1 (0)
	3	0 (0)	1 (0)	2 (0)	0 (0)	4 (.01)	2 (0)	2 (0)	3 (.01)	0 (0)	0 (0)
	2	0 (0)	2 (0)	9 (.02)	0 (0)	4 (.01)	10 (.02)	8 (.01)	2 (0)	0 (0)	0 (0)
	1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	1 (0)	0 (0)
	X	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	22 (.04)

Note. Cell values are raw numbers of sentences for USOs of Hispanic-Hispanic individuals at the indicated level of the Crime Seriousness Scale and the indicated level of the Criminal History Scale that were changed by the PHI adjustment followed, in parenthesis, by the proportion that number constitutes of the total number of adjusted sentences for USOs of Hispanic-Hispanic individuals. Light shading indicates percentage is between .05 and .09, inclusive; darker shading indicates percentages greater than .10. X indicates Unknown/Other.

Under circumstances like the sentencing decisions, where decisions are dichotomous and legitimate grounds for decision-making well specified, the concentration of adjustments to particular rating is likely not problematic for PHIs (although it may suggest specific targets for more effective preventative interventions; see e.g., McIntosh, Girvan, Fairbanks Falcon, et al, 2022). Where ratings are continuous, significant factors that raters are using to make decisions unknown or unincorporated into the PHI process, or both, however, contingent effects of rater' stereotypes and attitudes on r_0 may appear to be concentrated among certain rating ranges or otherwise non-linear. In such circumstances, efforts should be made to account for the moderating factors in the models used to calculate and apply adjustments.

Cure not Worse than the Disease

The final factor, embodied in certain anti-discrimination norms and legal doctrine, cautions generally against making direct adjustments to decision outcomes based on protected attributes of those involved as, itself, constituting discrimination. In effect, such adjustments are justified as a corrective measure only to the extent that we are sure that the group difference was caused by impermissible consideration of the attributes by raters (Chemerinsky, 2014; Girvan, 2020; Rutherglen, 2009). With respect to the application of PHIs in practice, whether the adjustment is an acceptable correction or unacceptable discrimination turns on the empirical strength of inference that the ratings were impacted by impermissible consideration of protected attributes as opposed to structural or other factors that happen to be correlated with those attributes. How strong the inference needs to be is itself a normative and legal question. As such, it could limit use of PHIs to the relatively narrow circumstances in which there is evidence of purposeful discrimination by the raters or extend PHIs to the relatively broad set of circumstances in which an objective observer could conclude from the available information that the protected attribute was a likely factor in the ratings (Sloan, 2020).

Application of PHIs to correct the sentencing decisions here likely falls between the two and perhaps satisfies both. There is no direct evidence that the judges who made the sentencing decisions did so in order to punish individuals that they perceived to be Hispanic more harshly, or, equivalently, those that they perceived as White more leniently. And I have made no effort to collect any. Even so, the combination of controls for the legally relevant information and finding that perceived, rather than self-identified race and ethnicity impacts sentencing outcomes for USOs on the margin is very consistent with

psychological theory regarding when raters' stereotypes and attitudes are most likely to impact their decisions. Accordingly, objective observers could certainly conclude that, consciously or unconsciously, the race and ethnicity of those being sentenced likely were a factor in the sentencing outcomes.

Conclusion

The goal of this paper is to use a sample of sentencing decisions to illuminate, explore, and examine the implications of a set of specified requirements for PHIs in practice. Among other things, the empirical thought experiment identified a common type of rating, dichotomous decisions such as whether an individual meets a certain threshold, that may be a challenging one in which to implement PHIs directly. In addition the example highlighted the potential importance of assessing stability in bias over time and modelling the specific nature of the biases in ratings, such as use of the same method to identify groups as did the raters. Finally, while, in practice, it may often be difficult to assess whether several of the requirements are strictly met, it may be possible to draw inferences about them. In those circumstances, the extent to which the inferences are sufficient to justify use of PHIs will likely turn on a normative and potentially legal question related to whether the correction itself is more problematic than its benefits.

Acknowledgements

This research was made possible through funding from the Pufendorf Institute for Advanced Studies at Lund University and the cooperation of the Oregon Criminal Justice Commission. The analysis and opinions expressed here are those of the author and do not necessarily represent the views of the Institute, Oregon Criminal Justice Commission, or the State of Oregon. Consistent with Open Data practices, a deidentified version of the data used for this study is available at https://osf.io/dr2wc. The author has no conflicts of interests to disclose.

References

Burch, T. (2015). Skin Color and the Criminal Justice System: Beyond Black-White Disparities in Sentencing. *Journal of Empirical Legal Studies*, *12*(3), 395-420.

Bushway, S. D., & Forst, B. (2013). Studying Discretion in the Processes that Generate Criminal Justice Sanctions. *Justice Quarterly*, *30*(2), 199-222.

Bushway, S. D., & Piehl, A. M. (2001). Judging Judicial Discretion: Legal Factors and Racial Discrimination in Sentencing. *Law and Society Review*, *35*(4), 733-764.

Chemerinsky, E. (2014). Making Schools More Separate and Unequal: Parents Involved in Community Schools v. Seattle School District No. 1. *Michigan State Law Review*, 633-665.

Crenshaw, K. W. (2017). *On intersectionality: Essential writings*. New York: The New Press.

Girvan, E. J. (2016). Wise Restraints?: Learning Legal Rules, Not Standards, Reduces the Effects of Stereotypes in Legal Decision-Making. *Psychology, Public Policy, and Law*, *22*(1), 31.

Girvan, E. J. (2020). Towards a Problem-Solving Approach to Addressing Racial Disparities in School Discipline Under Anti-Discrimination Law. *University of Memphis Law Review*, *50*, 995-1090.

Girvan, E. J. & Marek, H. (2023). Eye of the Beholder: Increased Likelihood of Prison Sentences for Those Perceived to Have Hispanic Ethnicity, *Law & Human Behavior* (in press).

Hassen, N., Lofters, A., Michael, S., Mall, A., Pinto, A. D., & Rackal, J. (2021). Implementing anti-racism interventions in healthcare settings: a scoping review. *International Journal of Environmental Research and Public Health*, *18*(6), 2993-3008.

Jönsson, M. L. and Bergman, J. (2022). Improving misrepresentations amid unwavering misrepresenters. *Synthese*, *200*.

Jönsson, M. L. and Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings. *Episteme*, *14*(4), 499 – 517.

King, R. D., & Light, M. T. (2019). Have Racial and Ethnic Disparities in Sentencing Declined?. *Crime and Justice*, *48*(1), 365-437.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., ... & Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative

investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., ... & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001-1016.

Liebman, J. S., Butler, K. C., & Buksunski, I. (2021). Mine the Gap: Using Racial Disparities to Expose and Eradicate Racism. *Southern California Review of Law & Social Justice*, *30*, 1-88.

Madva, A. (2020). Individual and Structural Interventions, in *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind* (eds. Beeghly, E. and Madva, A.). New York: Routledge.

McCall, L. (2005). The complexity of intersectionality. *Signs: Journal of Women in Culture and Society*, *30*(3), 1771-1800.

McIntosh, K., Girvan, E. J., Fairbanks Falcon, S., McDaniel, S. C., Smolkowski, K., Bastable, E., ... & Baldy, T. S. (2021). Equity-focused PBIS approach reduces racial inequities in school discipline: A randomized controlled trial. *School Psychology*, *36*(6), 433-444.

McIntosh, K., Smolkowski, K., Gion, C. M., Witherspoon, L., Bastable, E., & Girvan, E. J. (2020). Awareness is not enough: A double-blind randomized controlled trial of the effects of providing discipline disproportionality data reports to school administrators. *Educational Researcher*, *49*(7), 533-537.

Rees, H. R., Ma, D. S., & Sherman, J. W. (2020). Examining the Relationships Among Categorization, Stereotype Activation, and Stereotype Application. *Personality and Social Psychology Bulletin*, *46*(4), 499-513.

Rutherglen, G. (2009). Ricci v DeStefano: Affirmative action and the lessons of adversity. *The Supreme Court Review*, 2009(1), 83-114.

Sloan, A. (2020). "What to Do About Batson?": Using a Court Rule to Address Implicit Bias in Jury Selection, *California Law Review*, *108*, 233-266.