The Role of Beliefs in Driving Racial Discrimination: Experimental Evidence from the U.S. Pretrial System

Will Dobbie
Harvard & NBER

Crystal S. Yang
Harvard & NBER

March 2026

Motivation

- There is an increasingly widespread view that inaccurate beliefs are a key driver of discrimination, highlighting the need for information
  - Theoretical work shows how systematic biases can lead to incorrect stereotypes and discrimination (Bordalo et al. 2016, 2019)
  - Ban-the-Box policies reduce callbacks for Black applicants in audit studies, likely due to inaccurate stereotypes (Agan and Starr 2018)
  - Inaccurate beliefs drive disparities in online hiring experiments (Coffman et al. 2021), which disappear with information (Bohren et al. 2024)
- But it is unclear if findings will translate to the field or if frictions will limit the impact of information on beliefs and discriminatory behavior

The Bail System

- Following arrest, case/defendant information is presented to the bail judge who decides to effectively release or detain the individual
- Judges have a clear legal objective to release as many defendants as possible while minimizing failures to appear and new criminal activity
- Release decisions are usually made in just a few minutes, with limited feedback on release or misconduct rates after decisions are made
- High stakes for defendants, with lower conviction rates and higher employment rates if released before trial (Dobbie et al., 2018)
- Large White vs. Non-White release disparities, likely driven by limited information and/or inaccurate stereotypes (Arnold et al. 2018, 2022)

This Paper

- We study the results from two large randomized trials testing interventions to correct judges' beliefs and reduce release disparities
- First experiment provides judges with regular training and detailed court statistics, as well as individualized feedback every 3-4 months
  - Measure baseline and endline beliefs using online surveys and release and misconduct rates using administrative case data
  - Test whether we can correct the judges' beliefs and, by doing so, change release rates and reduce White vs. Non-White release disparities
  - Experiment provides first test of whether we can correct beliefs and reduce discrimination in field settings, not just the lab
- Second experiment tests external validity of our results using publicly-available data and a non-selected sample of judges in New York City

Preview of Findings

1. Judges have limited knowledge about their release and misconduct rates before the experiment, both overall and by race
2. Treated judges update their release and misconduct beliefs after receiving training, court statistics, and individualized feedback
3. But treated judges do not change their release rates or disparities, indicating that beliefs are not the sole driver of discrimination in bail

Preview of Findings

1. Judges have limited knowledge about their release and misconduct rates before the experiment, both overall and by race
   (i) Judges do not know their release rates and significantly overstate the likelihood that they release the same number of Whites and Non-Whites
   (ii) Judges also do not know their misconduct rates and significantly overstate the risk of releasing both Whites and Non-Whites
   (iii) The judges are internally consistent, with their White vs. Non-White release beliefs closely linked to their misconduct beliefs
2. Treated judges update their release and misconduct beliefs after receiving training, court statistics, and individualized feedback
3. But treated judges do not change their release rates or disparities, indicating that beliefs are not the sole driver of discrimination in bail

Preview of Findings

1. Judges have limited knowledge about their release and misconduct rates before the experiment, both overall and by race
2. Treated judges update their release and misconduct beliefs after receiving training, court statistics, and individualized feedback
   (i) Treated judges' release beliefs become more accurate and they become more likely to report White vs. Non-White release disparities
   (ii) Treated judges' misconduct beliefs also become more accurate, with more suggestive changes in White vs. Non-White misconduct disparities
3. But treated judges do not change their release rates or disparities, indicating that beliefs are not the sole driver of discrimination in bail

Preview of Findings

1. Judges have limited knowledge about their release and misconduct rates before the experiment, both overall and by race
2. Treated judges update their release and misconduct beliefs after receiving training, court statistics, and individualized feedback
3. But treated judges do not change their release rates or disparities, indicating that beliefs are not the sole driver of discrimination in bail
   (i) Statistically insignificant *decrease* in release and *increase* in White vs. Non-White release gap, both overall and for nearly all subsamples
   (ii) We confirm both results in our follow-up experiment in New York City using publicly-available data and a non-selected sample of judges
   (iii) Null results likely due to a combination of preference- and behaviorally-based frictions which limit the impact of beliefs on discrimination

Roadmap

1. Taking Theory to the Field
2. Baseline Beliefs
3. Treatment Effects on Endline Beliefs
4. Treatment Effects on Pretrial Outcomes
5. Mechanisms
6. Conclusion

The scandal of the criminal justice system is that we do what we do with no tracking of data or results.

– Judge from Wisconsin

Conceptual Framework

- Let $Y_i^* \in \{0,1\}$ indicate if defendant i would engage in misconduct if released and $D_{i,j} \in \{0,1\}$ indicate if defendant i is released by judge j
- Each judge uses available information such as race, $r_i \in \{b, w\}$, to form a potentially inaccurate prediction of misconduct risk:

$$h_{i,j} = \mathbb{E}^j[Y_i^* | r_i]$$

- Each judge has perceived costs of Type 1 ($T1$) and Type 2 errors ($T2$), which yield release preferences:

$$\tau_r^j = \frac{T2_r^j}{T1_r^j + T2_r^j}$$

Conceptual Framework

- Judges then make release decisions, which can be represented by an intended release threshold $D_{i,j}$:

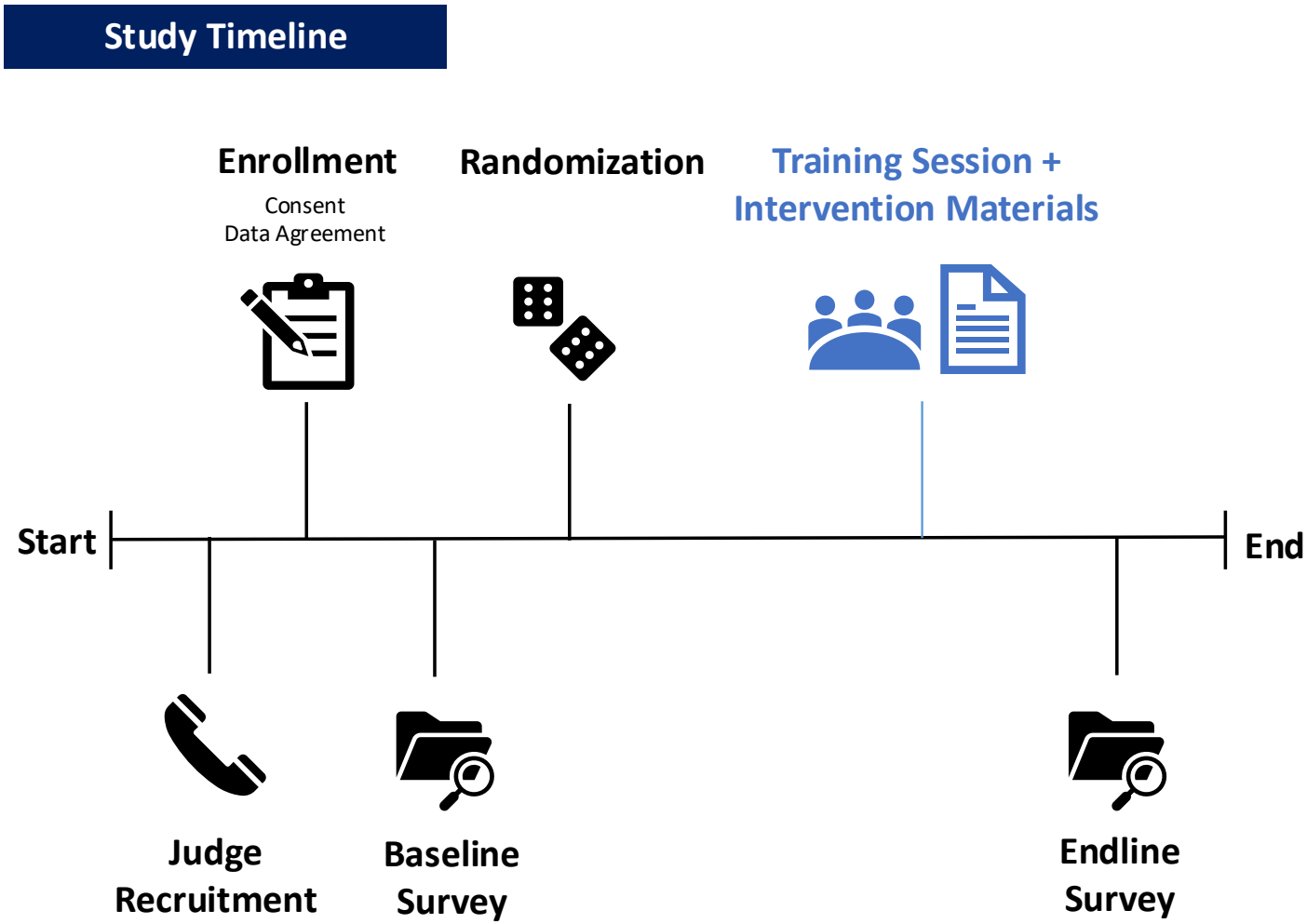
$$D_{i,j} = \mathbf{1}[h_{i,j} \leq \tau_r^j]$$

- With belief-driven disparities coming from inaccurate misconduct beliefs ($\mathbb{E}[h_{i,j}|r_i] \neq \mathbb{E}[Y_i^*|r_i]$) or release thresholds ($\mathbb{E}[D_{i,j}|r_i] \neq \tau_r^j$)

Conceptual Framework

- Under this standard framework, the ideal experiment to test whether we can correct inaccurate beliefs and reduce disparities is to:
  1. Elicit incentive-compatible beliefs on release and misconduct and compare to actual release and misconduct rates by defendant race
  2. Provide a random subset of judges with objective and easily understood information on release and misconduct rates by defendant race
  3. Estimate treatment effects on endline beliefs and actual release and misconduct rates by defendant race after the provision of information
- We bring this idealized experiment to the field using qualitative work, baseline and endline surveys, and a bundled information treatment

Experimental Design



Recruitment of Our Research Sites

- We recruited two court systems through extensive discussions with administrators, counsel, and the lead judges (when relevant)
- Judges were then individually consented, with an NIH Certificate of Confidentiality ensuring confidentiality ▸ Consent ▸ Certificate of Confidentiality
- We also took steps to increase buy-in by providing private feedback to the judges, with materials tailored to each court system

Our First Research Site

- Our first research site is a court in a large mid-Atlantic city that handles both felony and misdemeanor cases
  - Successfully recruited 35 of 43 full-time judges at baseline and 2 of 8 judges who joined the court after the project started ▸ Representativeness
  - Judges handled an average of 345 cases in the two years before the experiment and 371 cases during the two-year project
  - Judges are elected via competitive elections to serve six-year terms, with the participating judges being 78% male and 89% White
  - White defendants 5.2pp more likely to be released than Non-White defendants before the experiment

Our Second Research Site

- Our second research site is a court in a large west coast city that only handles misdemeanor cases
  - Successfully recruited 12 of 12 full-time judges at baseline and 2 of 2 judges who joined the court after the project started
  - Judges handled an average of 380 cases in the two years before the experiment and 368 cases during the two-year project
  - Judges are court-appointed or elected via competitive elections to serve four-year terms, with the judges being 57% male and 35% White
  - White defendants 4.9pp more likely to be released than Non-White defendants before the experiment

Taking Theory to the Field

- There are several important ways that real-world interventions in the pretrial system will differ from the idealized lab equivalent:
  - Judges are professional and motivated public servants, with many facing election and political risk that are not present in lab settings
  - Decision-making process more complex and outcomes observed over longer periods than stylized one-shot decisions in lab settings
  - Misconduct is only observed among released individuals, making it difficult to provide feedback on unconditional risk as in lab settings
  - Judges view financial incentives as a method of eliciting accurate survey responses to be disrespectful, consistent with Harrington et al. (2025)
- We worked with 20+ judges to develop surveys and materials that allow us to take the theory to the field given these differences

Baseline Survey

- Judges sent a baseline survey before randomization, with 48 out of the 51 participating judges taking the survey
- Elicited non-incentivized beliefs about release and conditional misconduct for all, White, and Non-White defendants ▸ Examples
- Elicited preferences with hypothetical release increases to compensate for fixed misconduct increase, yielding "price" for misconduct ▸ Example
- We also administered an IAT test and asked about peer practices to elicit potentially more truthful/implicit beliefs about bias ▸ Examples

Baseline Outcomes

- We match survey responses to case data spanning up to two years before the experiment with objective release and misconduct rates
- Observe defendant demographics, offense, criminal history, release decision, misconduct (FTA or NCA) if released, and the judge
- Following our pre-analysis plan, we weight results on baseline beliefs by the number of cases to account for noise from small caseloads
- Final sample for baseline analysis includes $\approx$ 18k cases, $\approx$ 14k defendants, and 51 judges ▸ Summary Statistics

Bundled Information Treatment

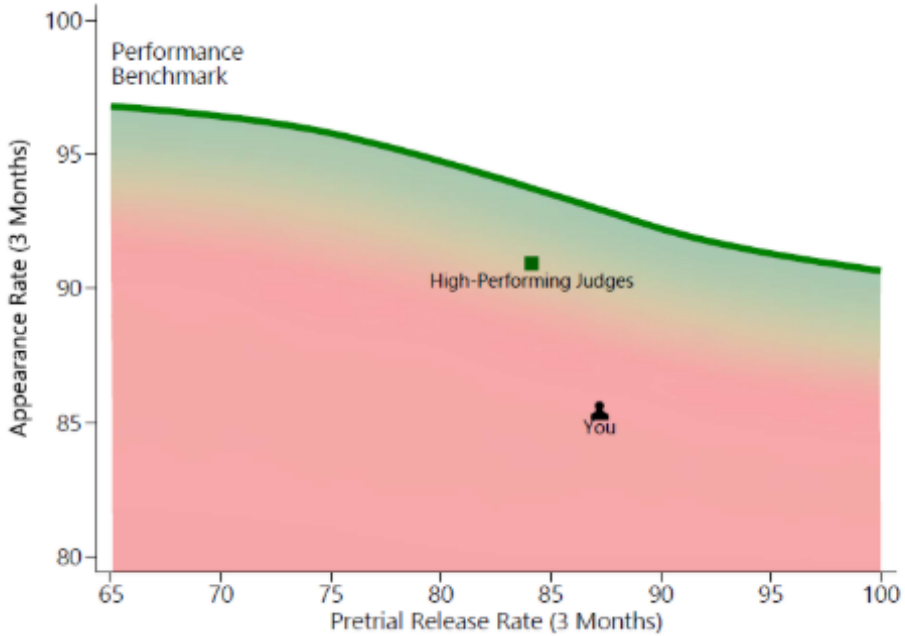
- We provide a random subset of judges with a bundled intervention treatment to correct mistaken beliefs about release and misconduct:
  1. Individualized feedback about judges' release and conditional FTA and NCA rates by defendant race, e.g., $\mathbb{E}[D_{i,j}|r_i]$ and $\mathbb{E}[Y_i^*|D_{i,j}=1, r_i]$
  2. Court statistics aggregated across all judges on release and FTA and NCA rates by defendant race, e.g., $\mathbb{E}[Y_i^*|r_i]$
- We provided treated judges with feedback every 3-4 months for two years, with 96% take-up based on round-by-round confirmations
- We also held training sessions at the start of the project and offered to meet with the judges after each round to interpret the materials

Individualized Feedback Example

Your performance statistics for October 1, 2022 – February 28, 2023

	All Defendants	White Defendants	Non-White Defendants
Pretrial Release Rate (3 Months)	87%	92%	85%
Appearance Rate (3 Months)	85%	88%	84%
Arrest-Free Rate (3 Months)	92%	92%	92%

Individualized Feedback Example



HOW YOU COMPARE

	Great
	Good
X	Average/Below Average

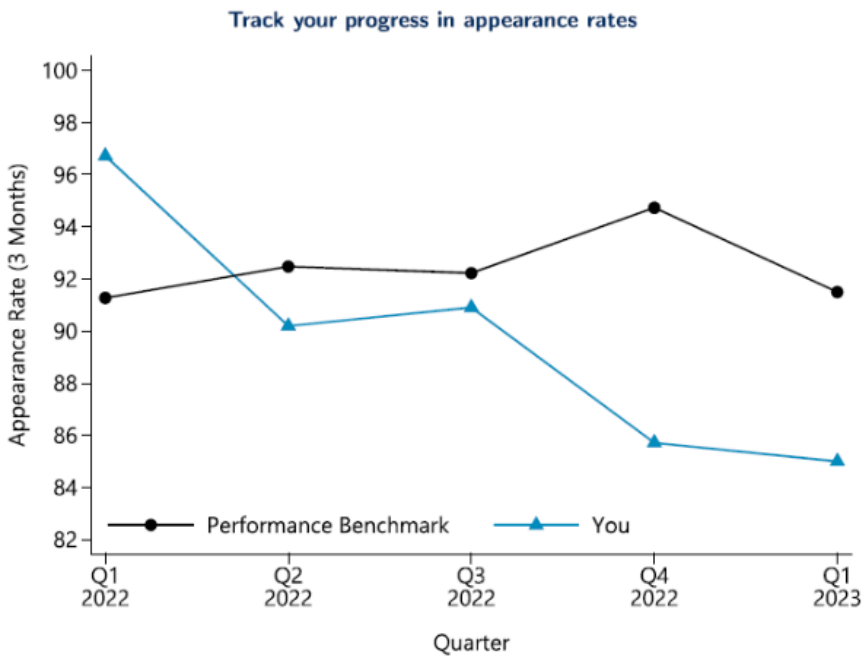
Your release rate is 87%, 13% worse than the performance benchmark.

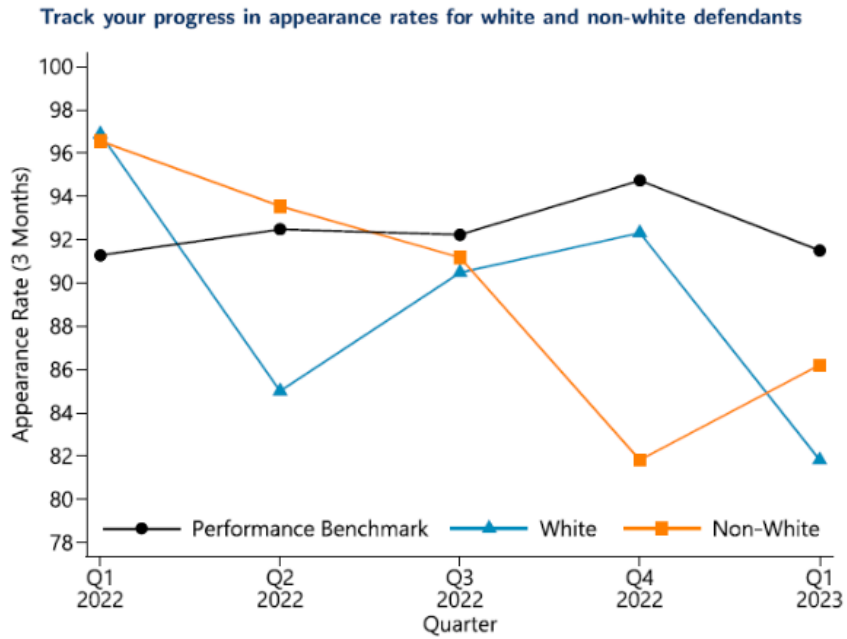
Your appearance rate is 85%, 8% worse than the performance benchmark.

*The performance benchmark is the best possible combination of pretrial release rates and appearance rates that can be achieved if decisions were made using all available data on defendants.

**High-performing judges are the judges in your court with the highest pretrial release rates and highest appearance rates among released defendants.

Individualized Feedback Example



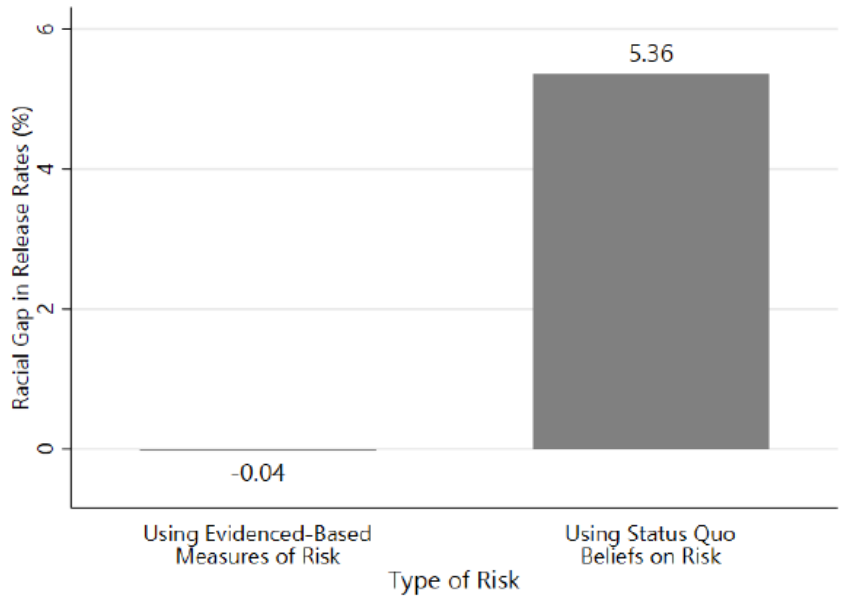


Court Statistics Example

Correcting Mistaken Beliefs on Defendant Risk

Measuring Risk as Rearrest– All Cases

Here's how the white/non-white racial gap in release rates differs if release decisions were made using *evidence-based measures of risk* vs. *status quo beliefs on risk.*



KEY TAKEAWAYS

- The racial gap in release rates should only be -0.04% when using evidence-based measures of risk, but the racial gap is 5.36% when using status quo beliefs on risk
- This implies that the racial gap in release rates is not justified using evidence-based measures of risk and that non-white defendants should be released at much higher rates

Endline Survey

- Endline survey sent $\approx$ two years after random assignment, with 40 out of 51 judges taking the survey, balanced by treatment status
- Similar to baseline, we did not incentivize responses and we allowed empty responses if judges were not able to guess an answer
- 85% of treated judges reported using intervention materials in the survey, with 70% wanting continued feedback ▸ Endline Feedback
- Endline survey matched to case data in the two years after randomization, with $\approx$ 28k cases, $\approx$ 21k defendants, and 51 judges

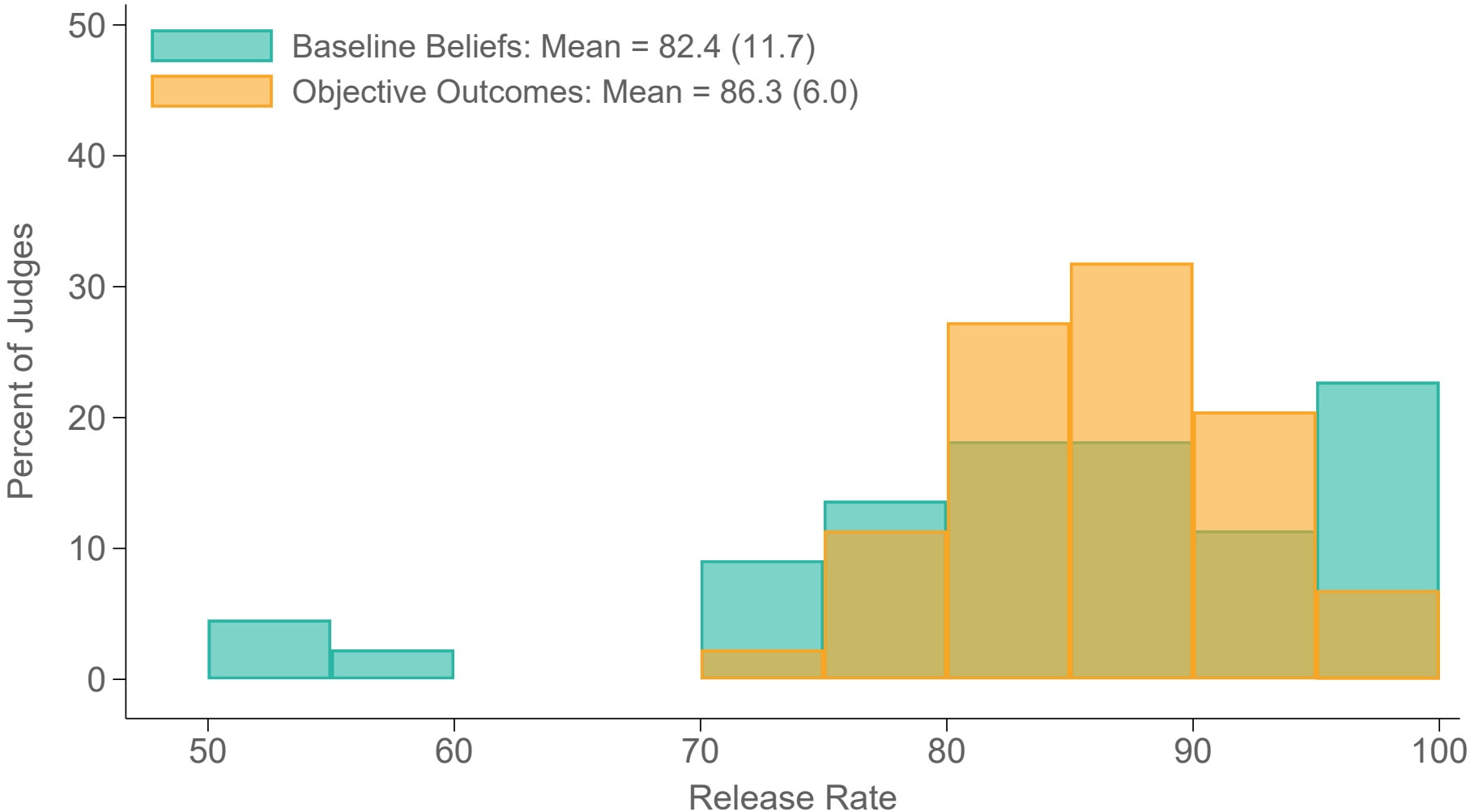
Experimenter Demand Effects

- The endline survey was designed to minimize experimenter demand effects as much as possible given the study design and setting
  - Endline survey sent over two years after the start of the project
  - Endline survey administered primarily online
  - Beliefs questions come before questions about race and the IAT
- Release and misconduct questions also do not have a preferred answer from our perspective, further minimizing experimenter demand effects

Roadmap

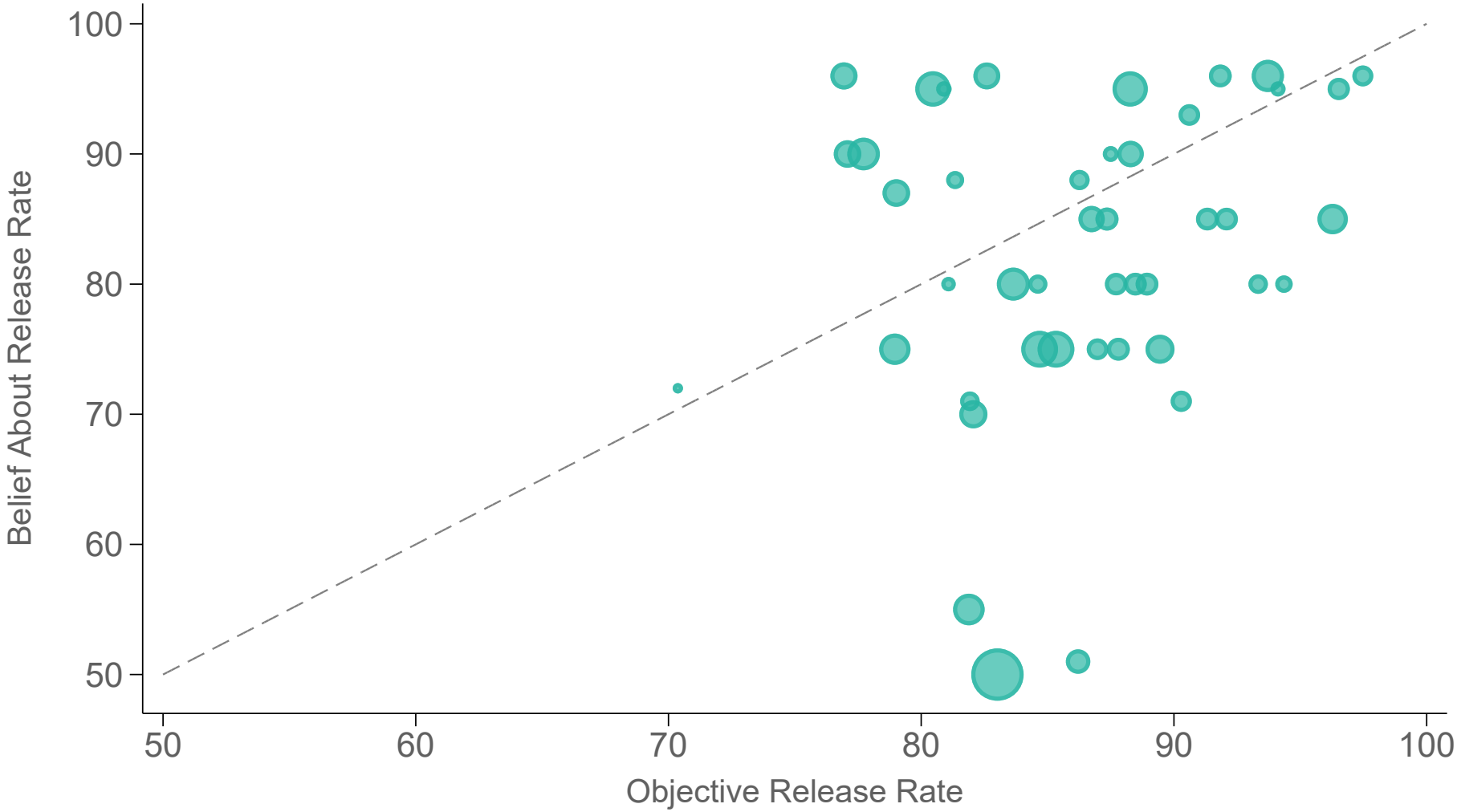
1. Taking Theory to the Field
2. Baseline Beliefs
3. Treatment Effects on Endline Beliefs
4. Treatment Effects on Pretrial Outcomes
5. Mechanisms
6. Conclusion

Baseline Release Beliefs vs. Objective Release Outcomes



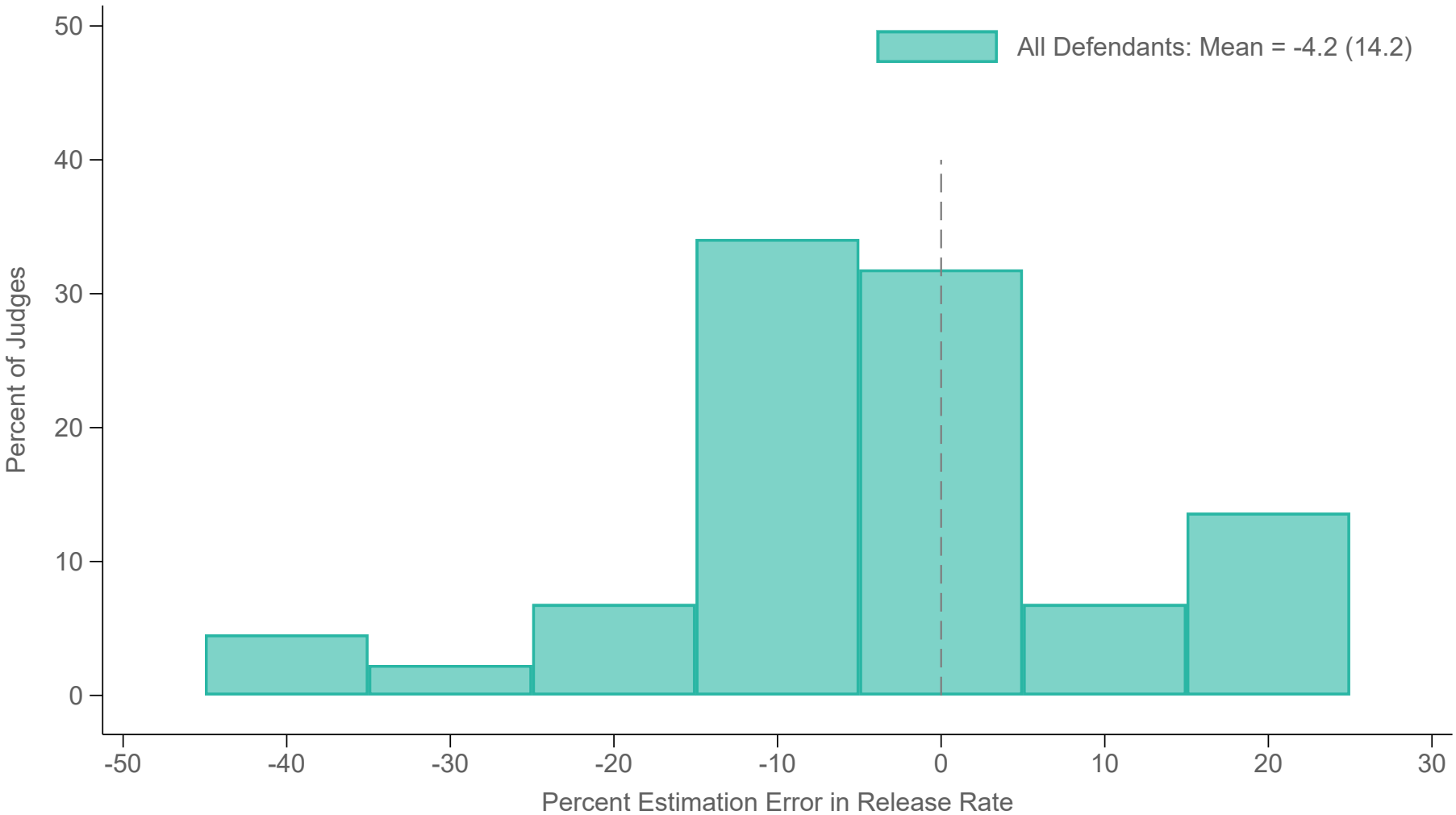
Note. This figure plots baseline release beliefs and objective release outcomes for all defendants and judges (N=44). Four judges coded as "I don't know."

Baseline Release Beliefs vs. Objective Release Outcomes



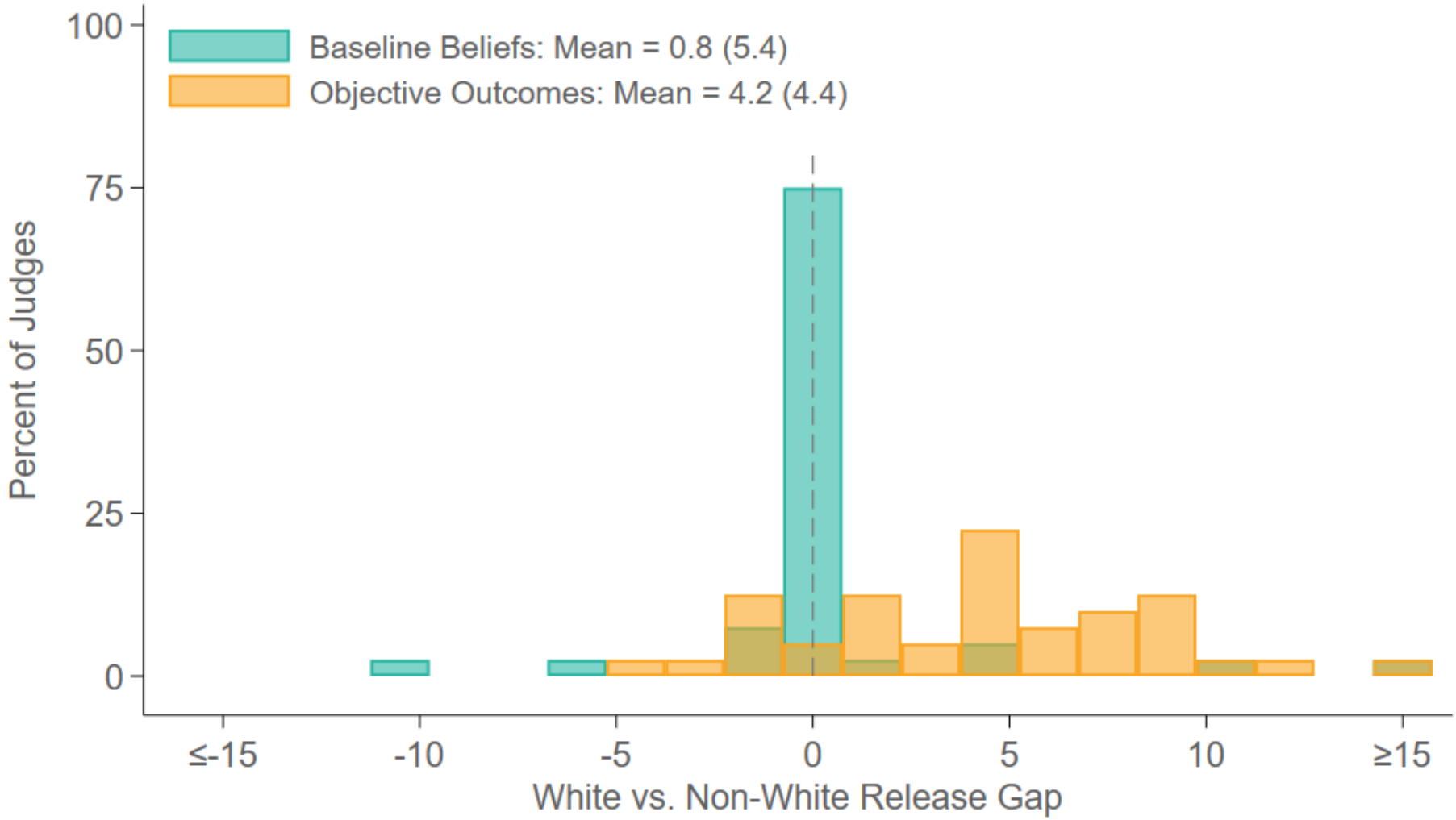
Note. This figure plots baseline beliefs and objective release outcomes (N=44). Four judges coded as "I don't know." Dashed line indicates accurate beliefs with constant=0, slope=1.

Estimation Error for Baseline Release Beliefs



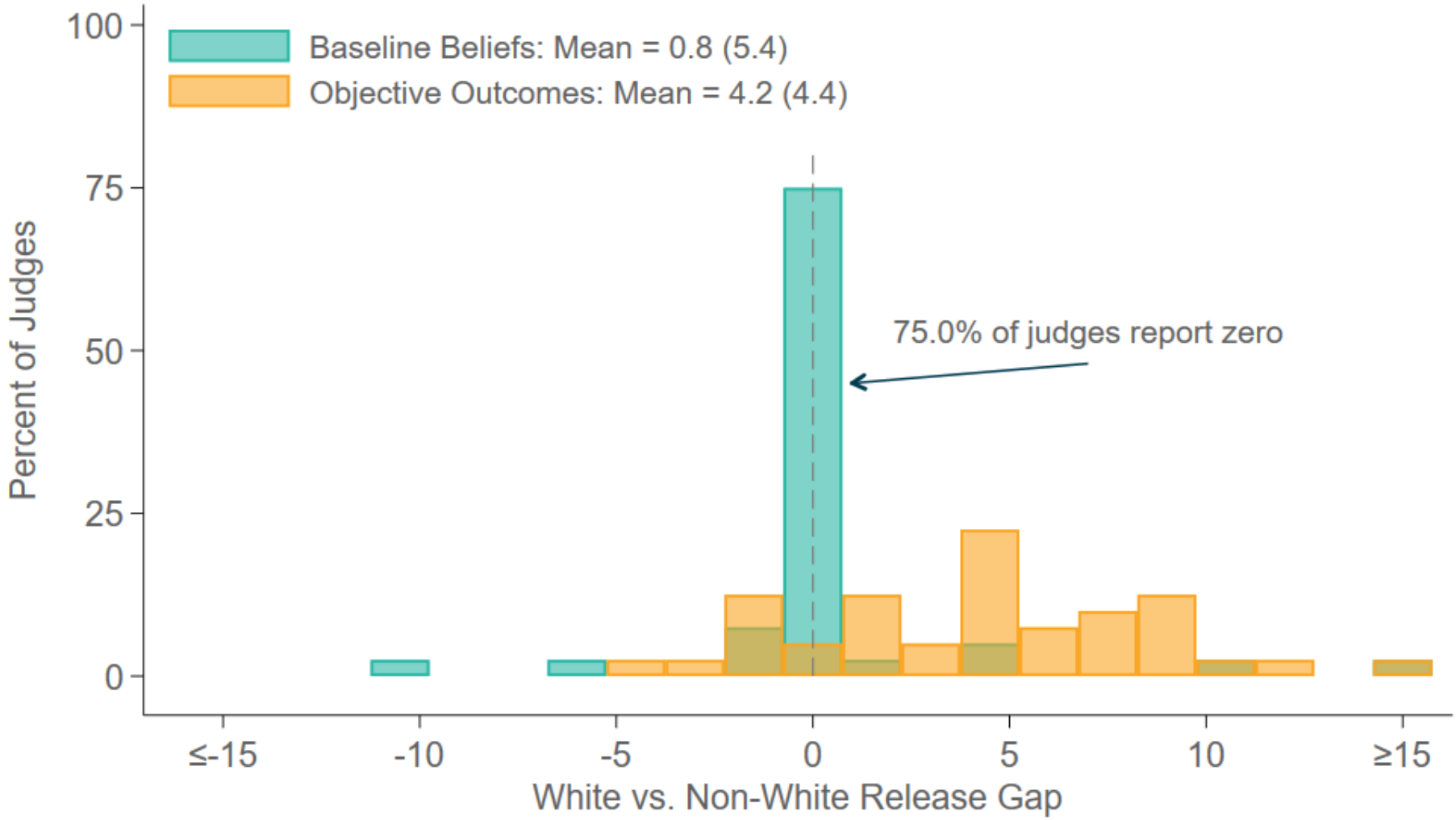
Note. This figure plots the percentage difference between baseline release beliefs and objective release outcomes (N=44). Four judges coded as "I don't know." ▸ Estimation Error by Race

Baseline White vs. Non-White Release Beliefs



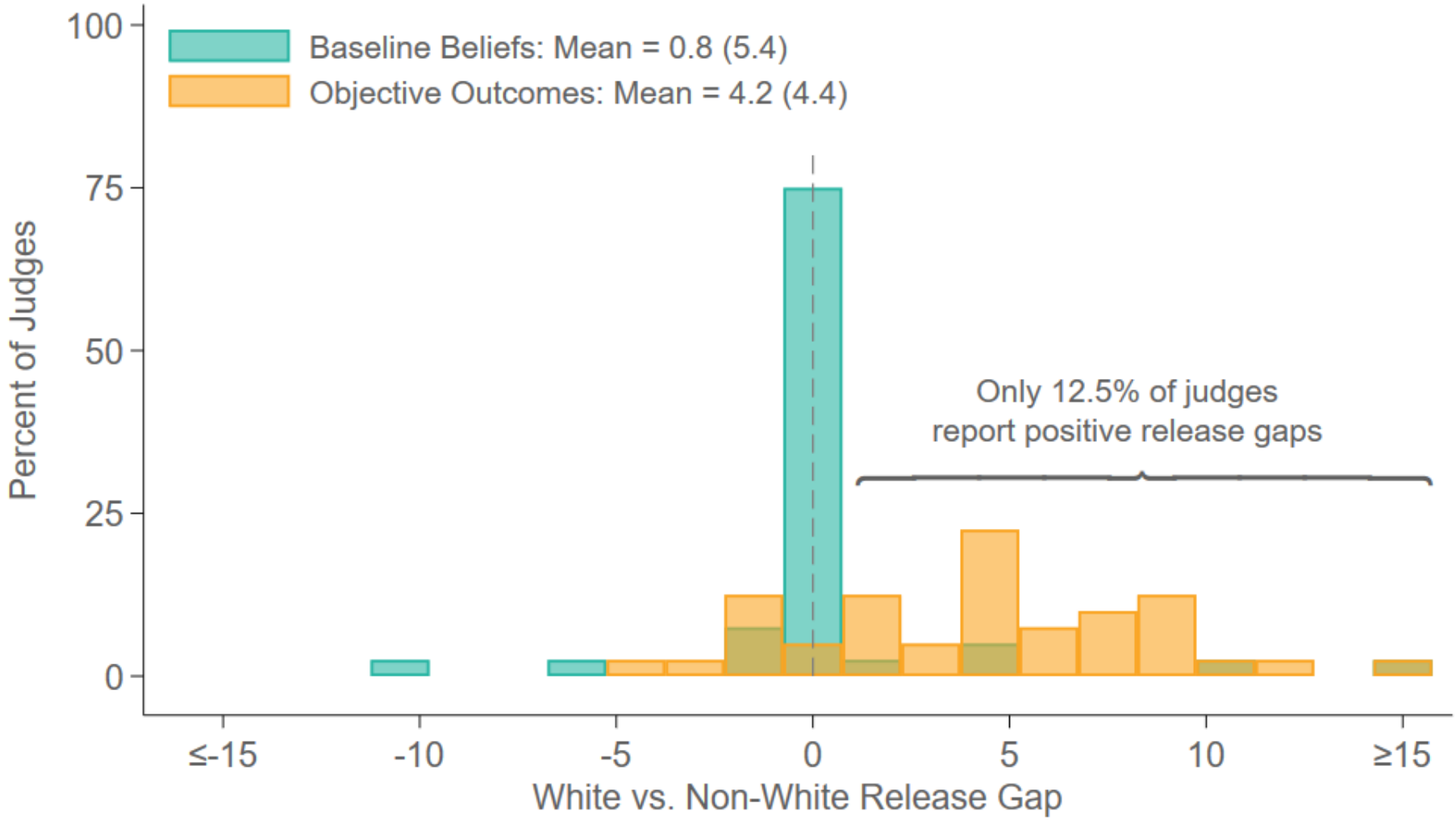
Note. This figure plots baseline White vs. Non-White release beliefs and objective White vs. Won-White release disparities (N=40). Eight judges coded as "I don't know."

Baseline White vs. Non-White Release Beliefs



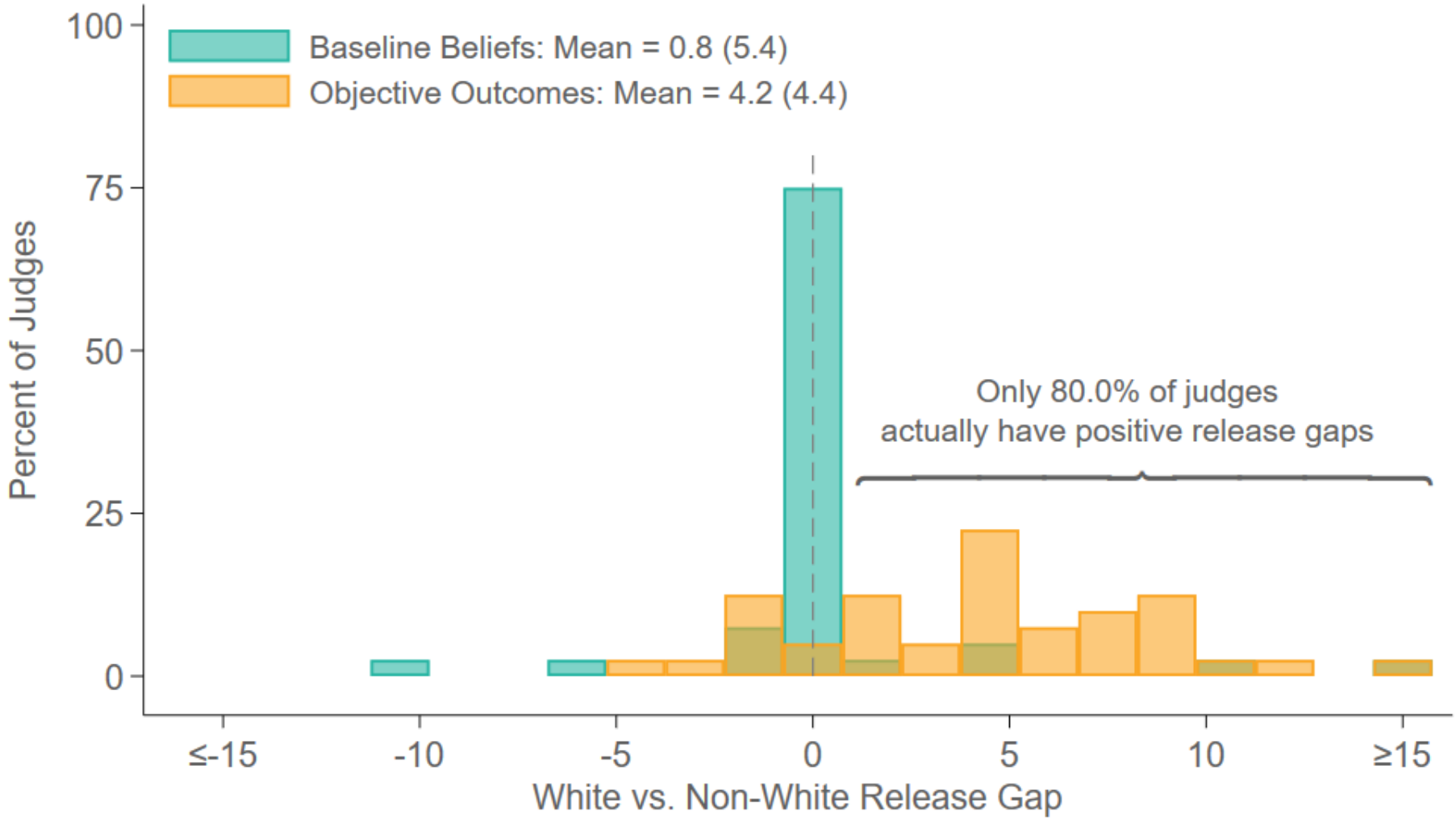
Note. This figure plots baseline White vs. Non-White release beliefs and objective White vs. Non-White release disparities (N=40). Eight judges coded as "I don't know."

Baseline White vs. Non-White Release Beliefs



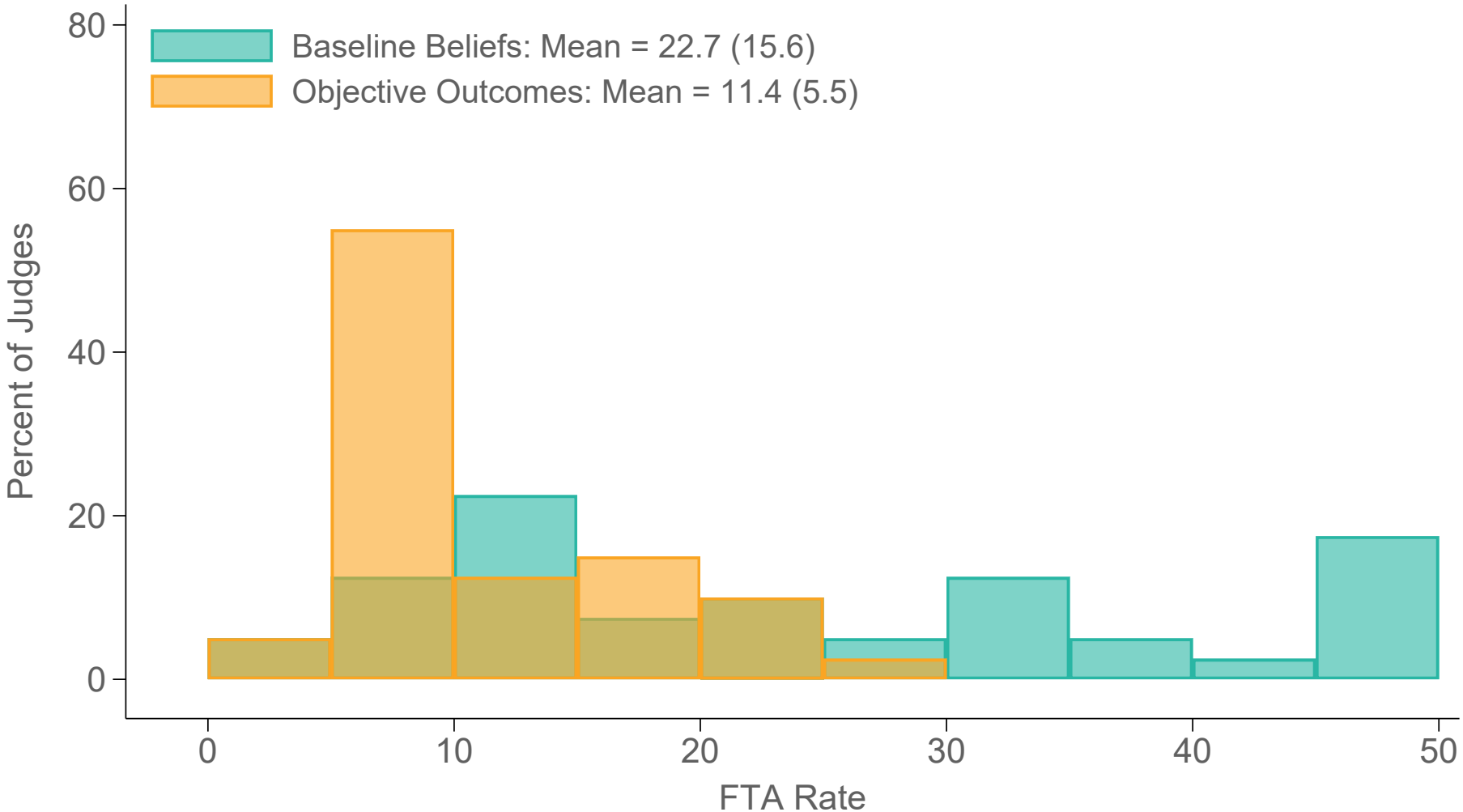
Note. This figure plots baseline White vs. Non-White release beliefs and objective White vs. Non-White release disparities (N=40). Eight judges coded as "I don't know."

Baseline White vs. Non-White Release Beliefs



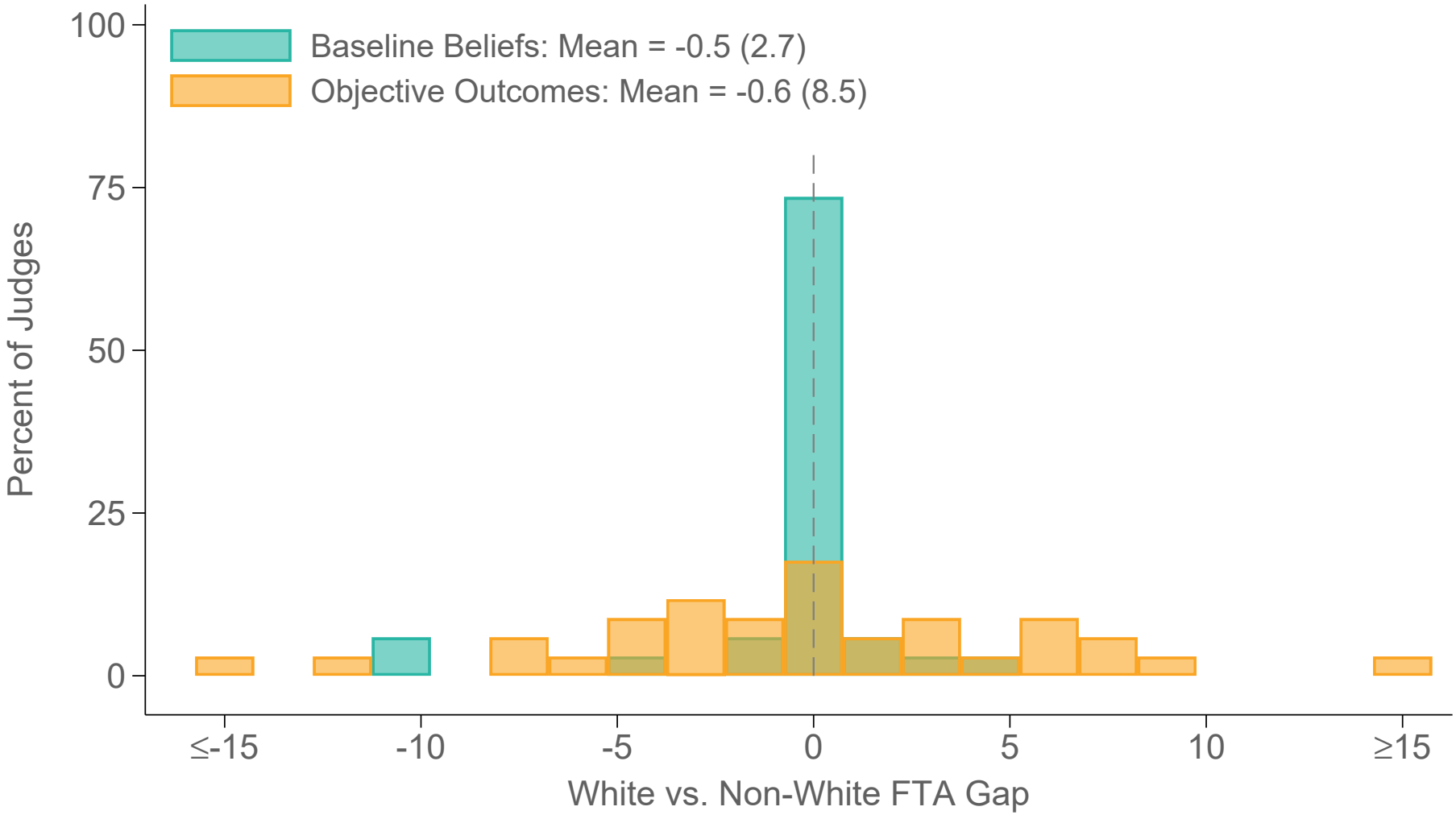
Note. This figure plots baseline White vs. Non-White release beliefs and objective White vs. Non-White release disparities (N=40). Eight judges coded as "I don't know."

Baseline FTA Beliefs vs. Objective FTA Outcomes



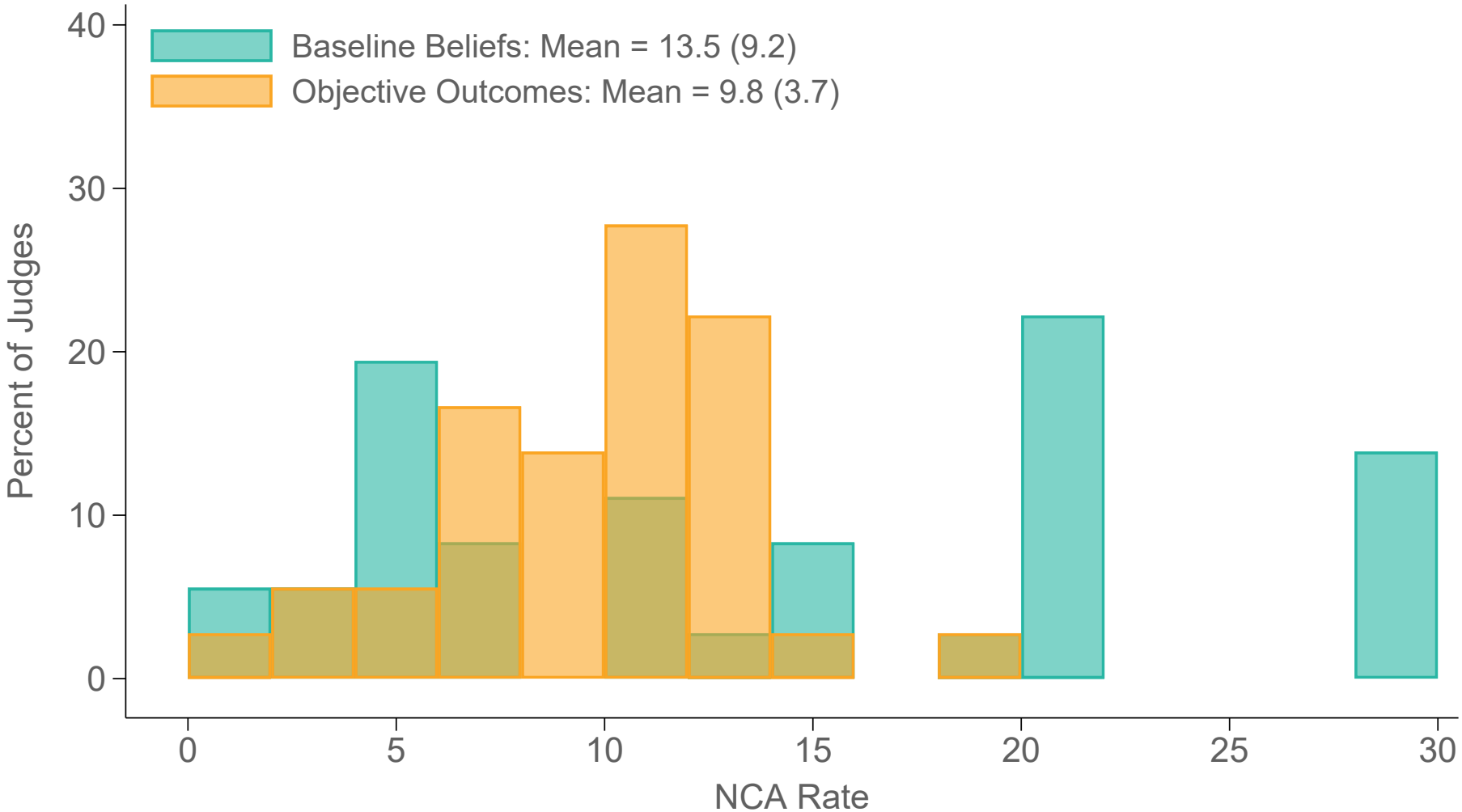
Note. This figure plots baseline FTA beliefs and objective FTA outcomes for all defendants and judges (N=40). Eight judges coded as "I don't know."

Baseline White vs. Non-White FTA Beliefs



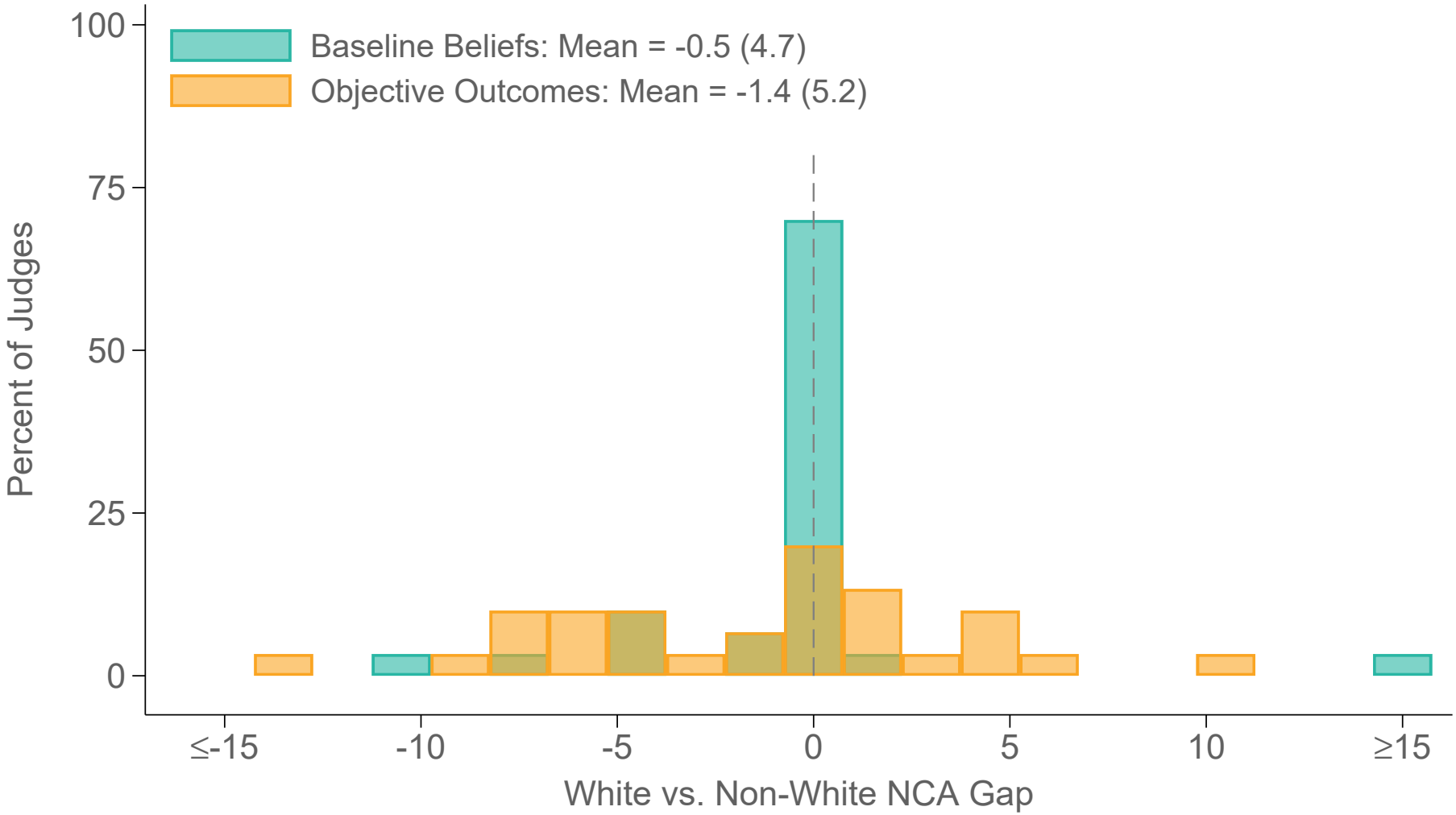
Note. This figure plots baseline White vs. Non-White FTA beliefs and objective White vs. Non-White FTA disparities (N=34). Fourteen judges coded as "I don't know."

Baseline NCA Beliefs vs. Objective NCA Outcomes



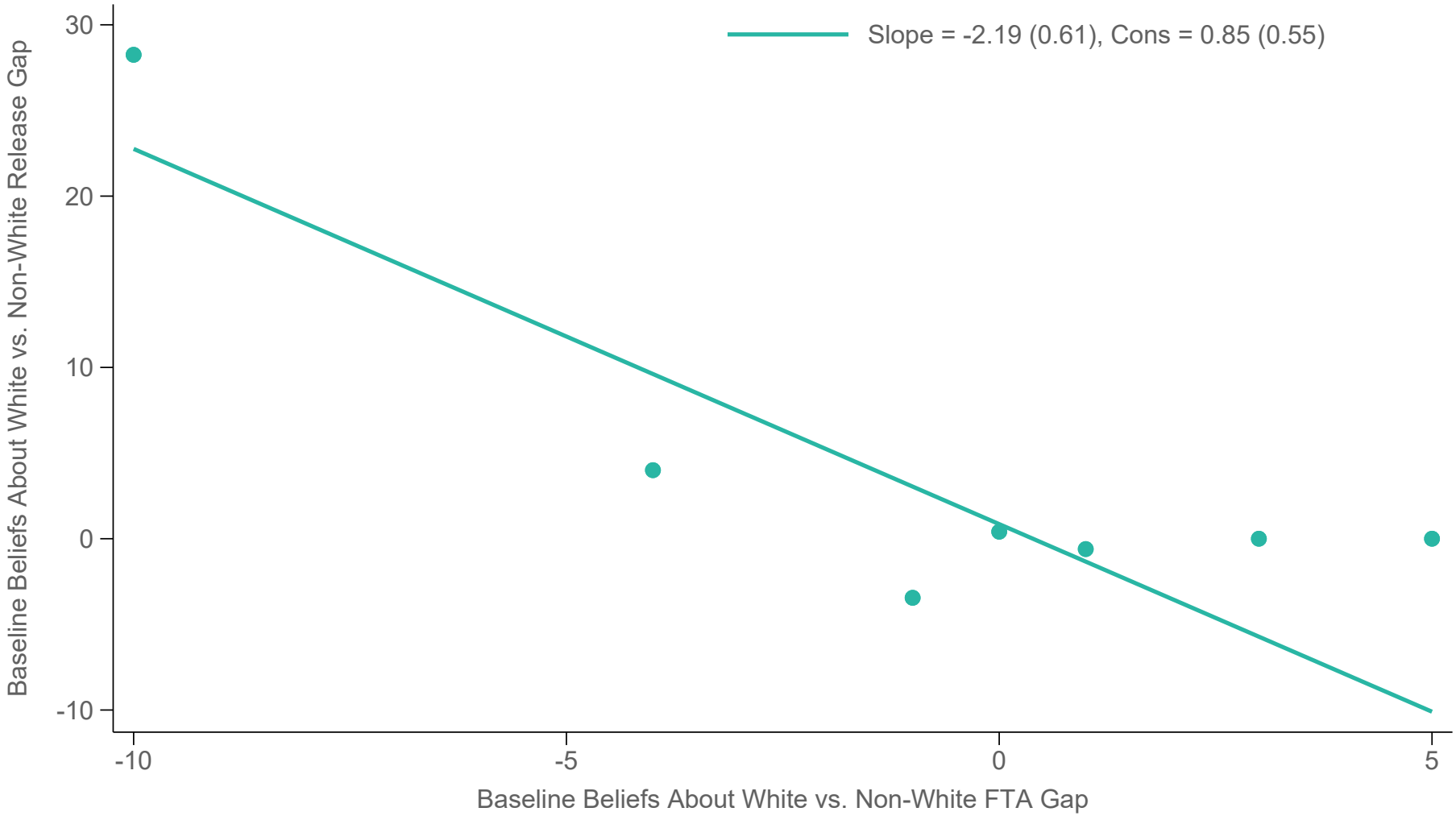
Note. This figure plots baseline NCA beliefs and objective NCA outcomes for all defendants and judges (N=36). Twelve judges coded as "I don't know."

Baseline White vs. Non-White NCA Beliefs



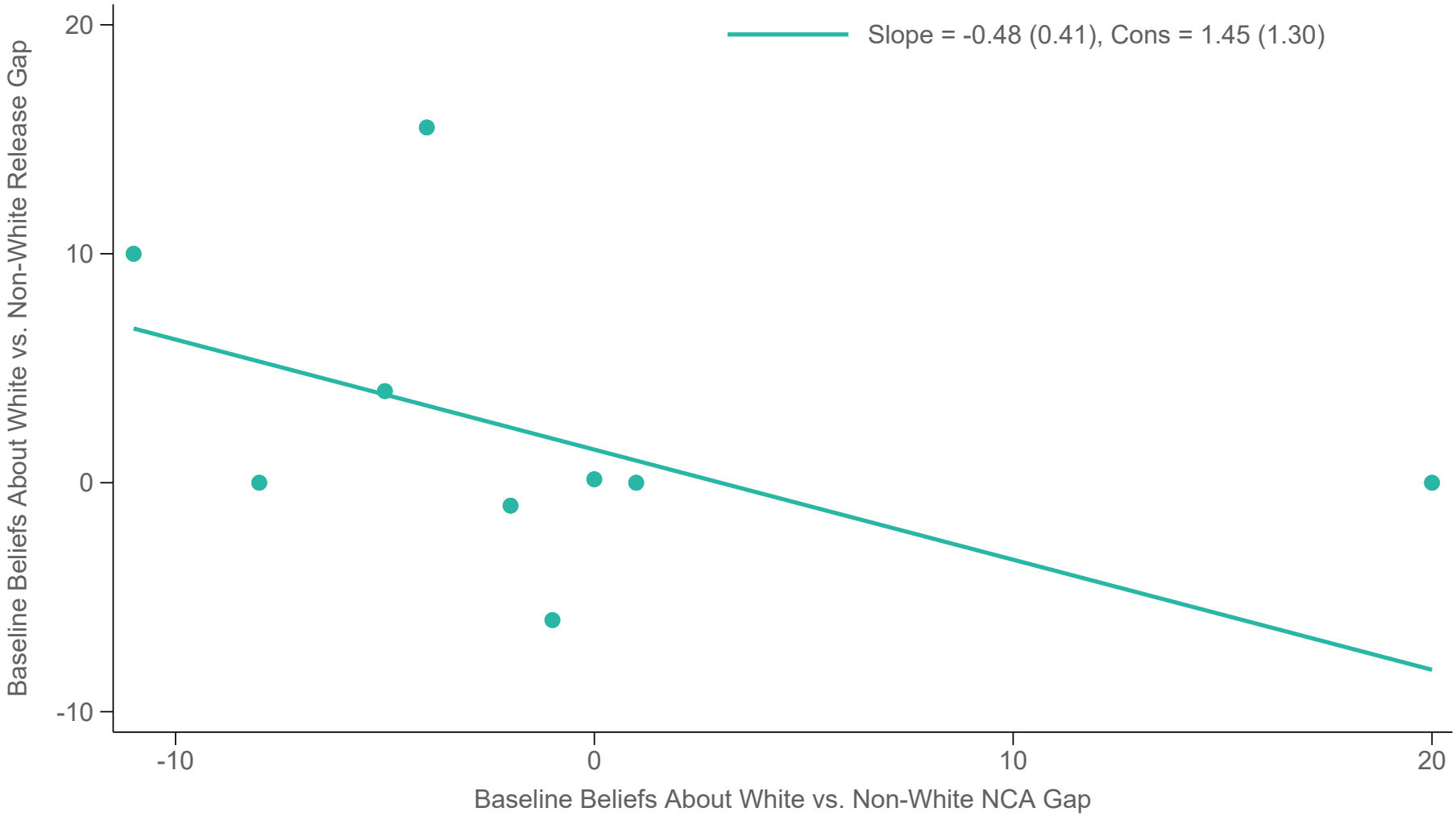
Note. This figure plots baseline White vs. Non-White NCA beliefs and objective White vs. Non-White NCA disparities (N=30). Eighteen judges coded as "I don't know."

Baseline Release vs. FTA Beliefs



Note. This figure plots a binscatter of baseline release beliefs vs. baseline FTA beliefs (N=34). Fourteen judges coded as "I don't know."

Baseline Release vs. NCA Beliefs

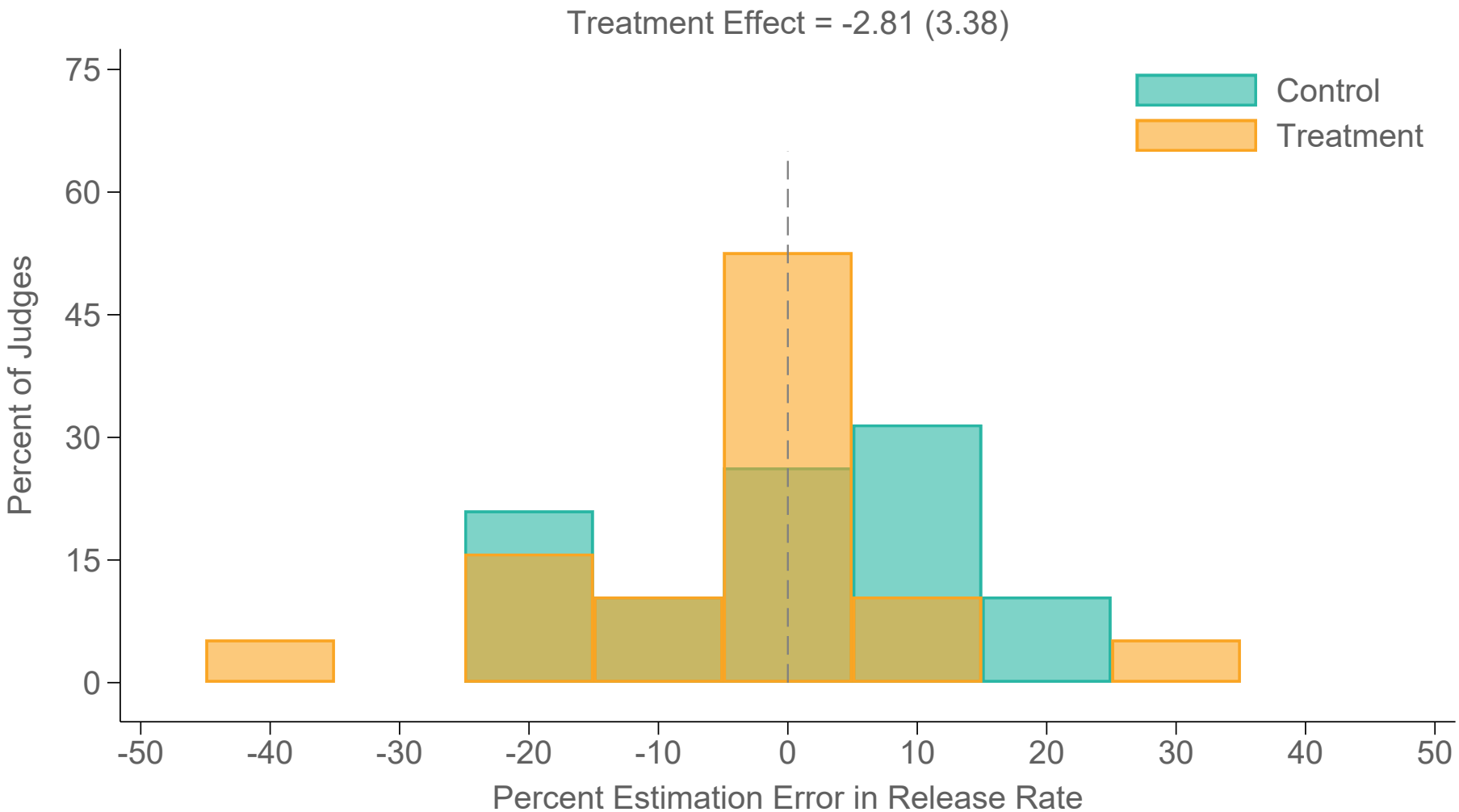


Note. This figure plots a binscatter of baseline release beliefs vs. baseline NCA beliefs (N=30). Eighteen judges coded as "I don't know."

Roadmap

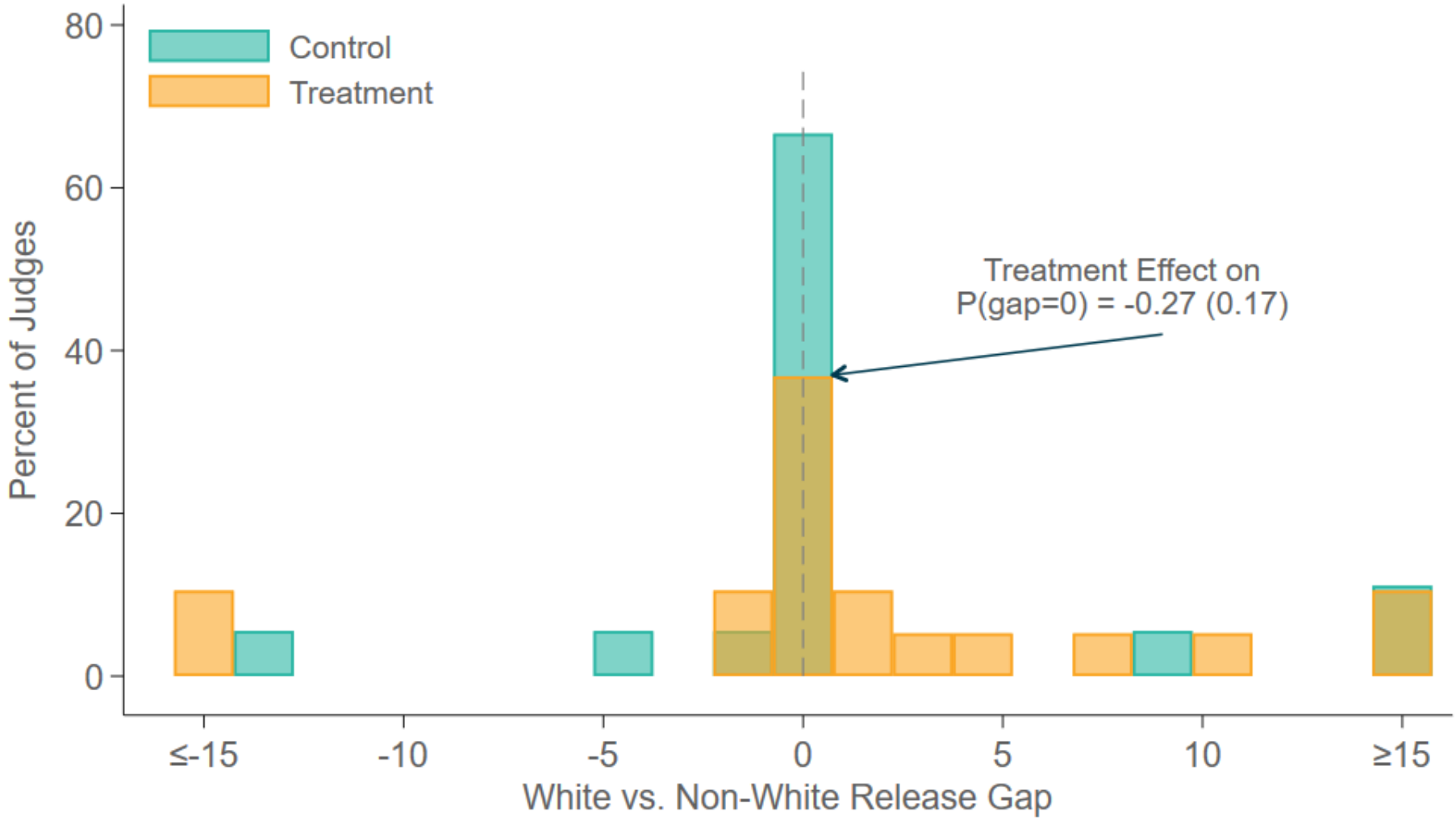
1. Taking Theory to the Field
2. Baseline Beliefs
3. Treatment Effects on Endline Beliefs
4. Treatment Effects on Pretrial Outcomes
5. Mechanisms
6. Conclusion

Estimation Error for Endline Release Beliefs



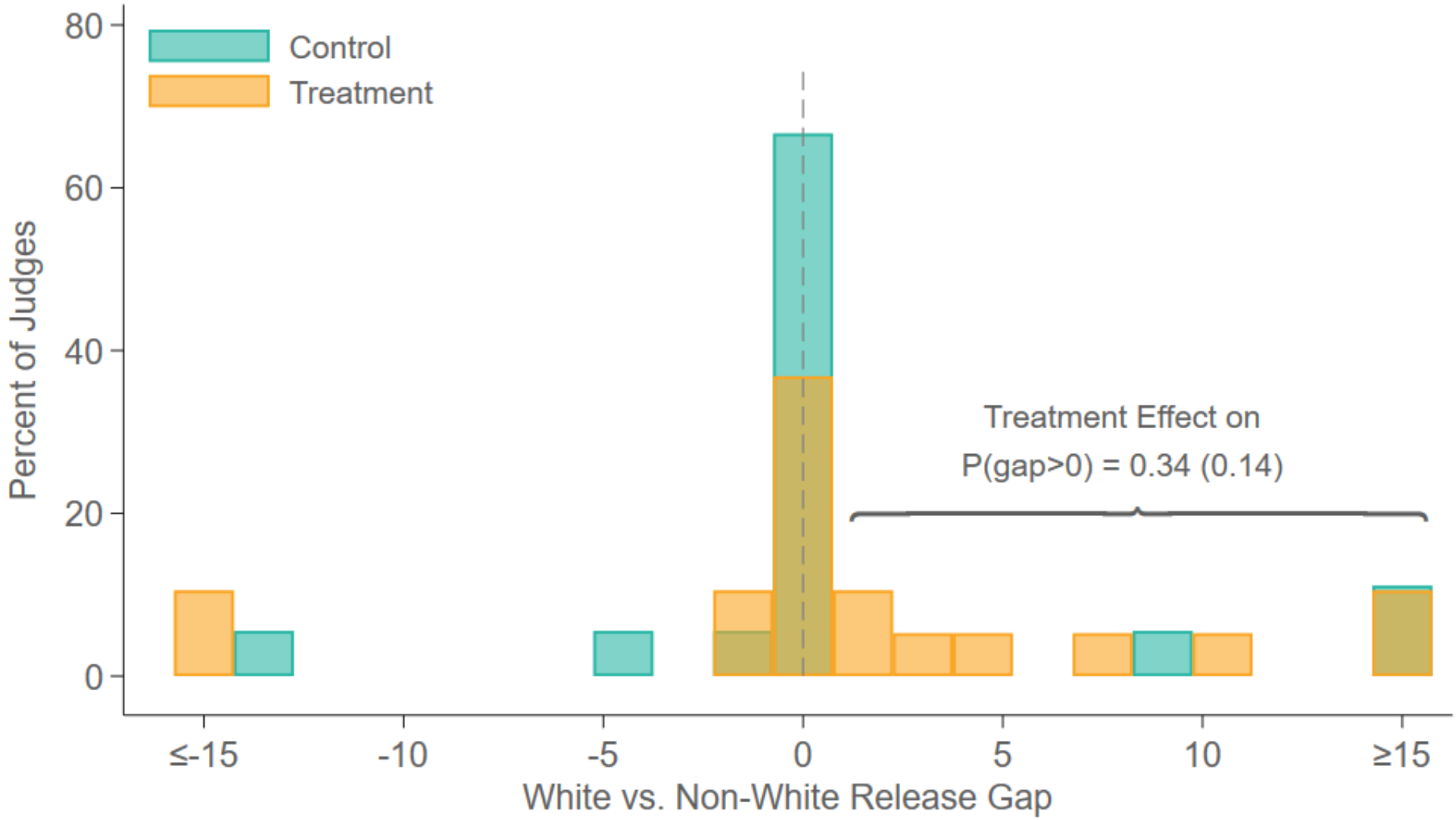
Note. This figure plots control vs. treatment judges' endline beliefs for all defendants (N=38). Two judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects. ▸ White ▸ Non-White

Endline White vs. Non-White Release Beliefs



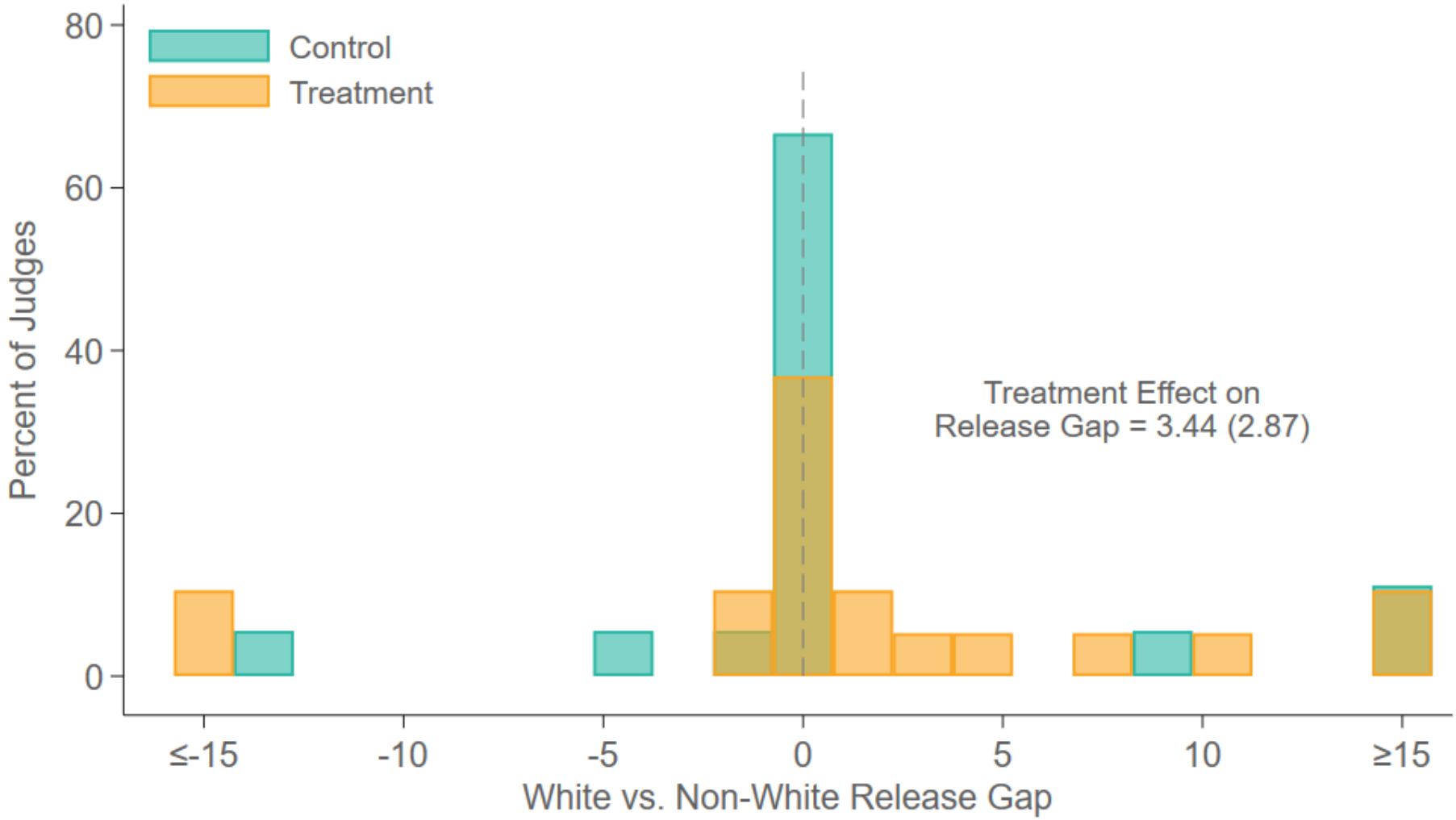
Note. This figure plots control vs. treatment judges' endline beliefs by race (N=37). Three judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects.

Endline White vs. Non-White Release Beliefs



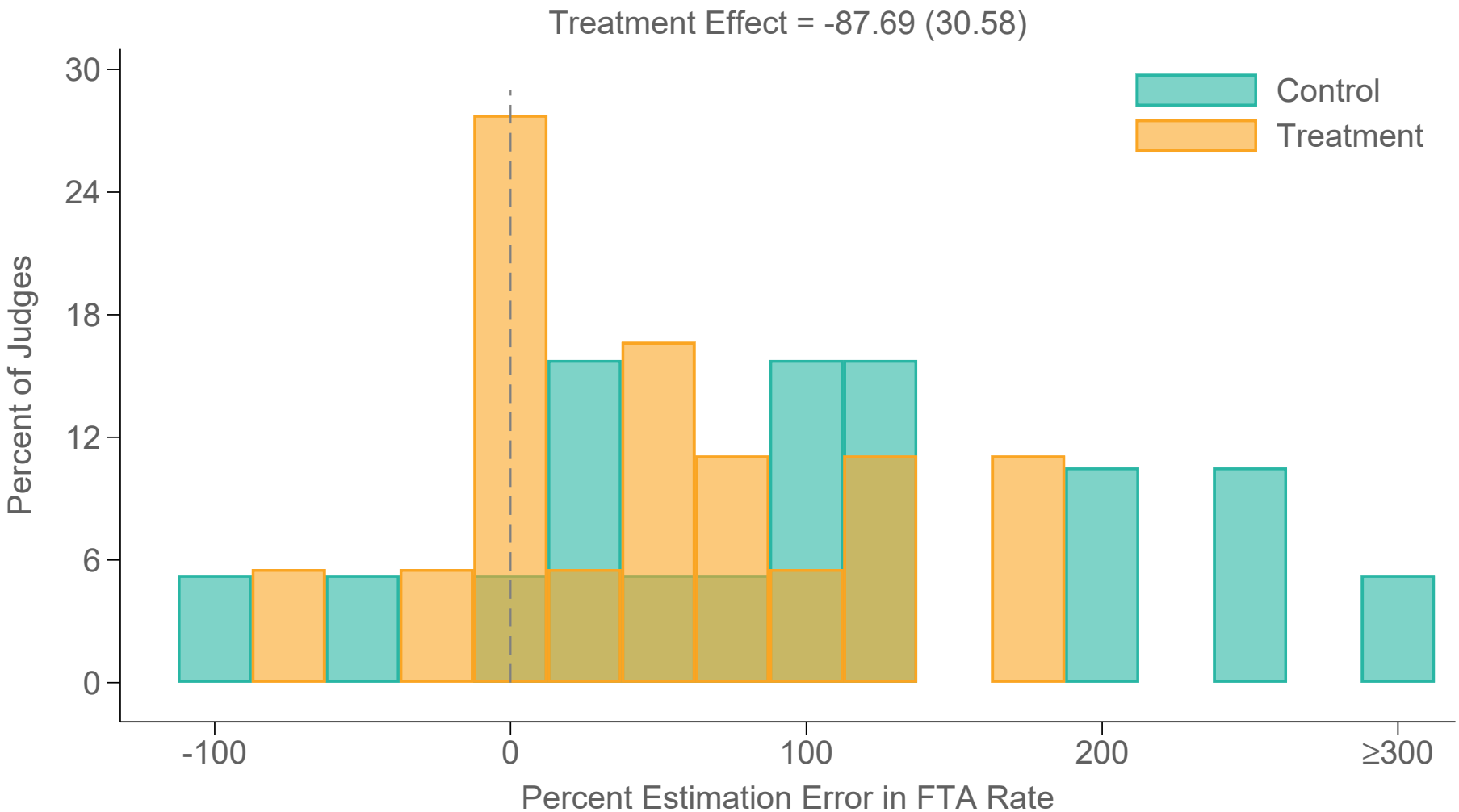
Note. This figure plots control vs. treatment judges' endline beliefs by race (N=37). Three judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects.

Endline White vs. Non-White Release Beliefs



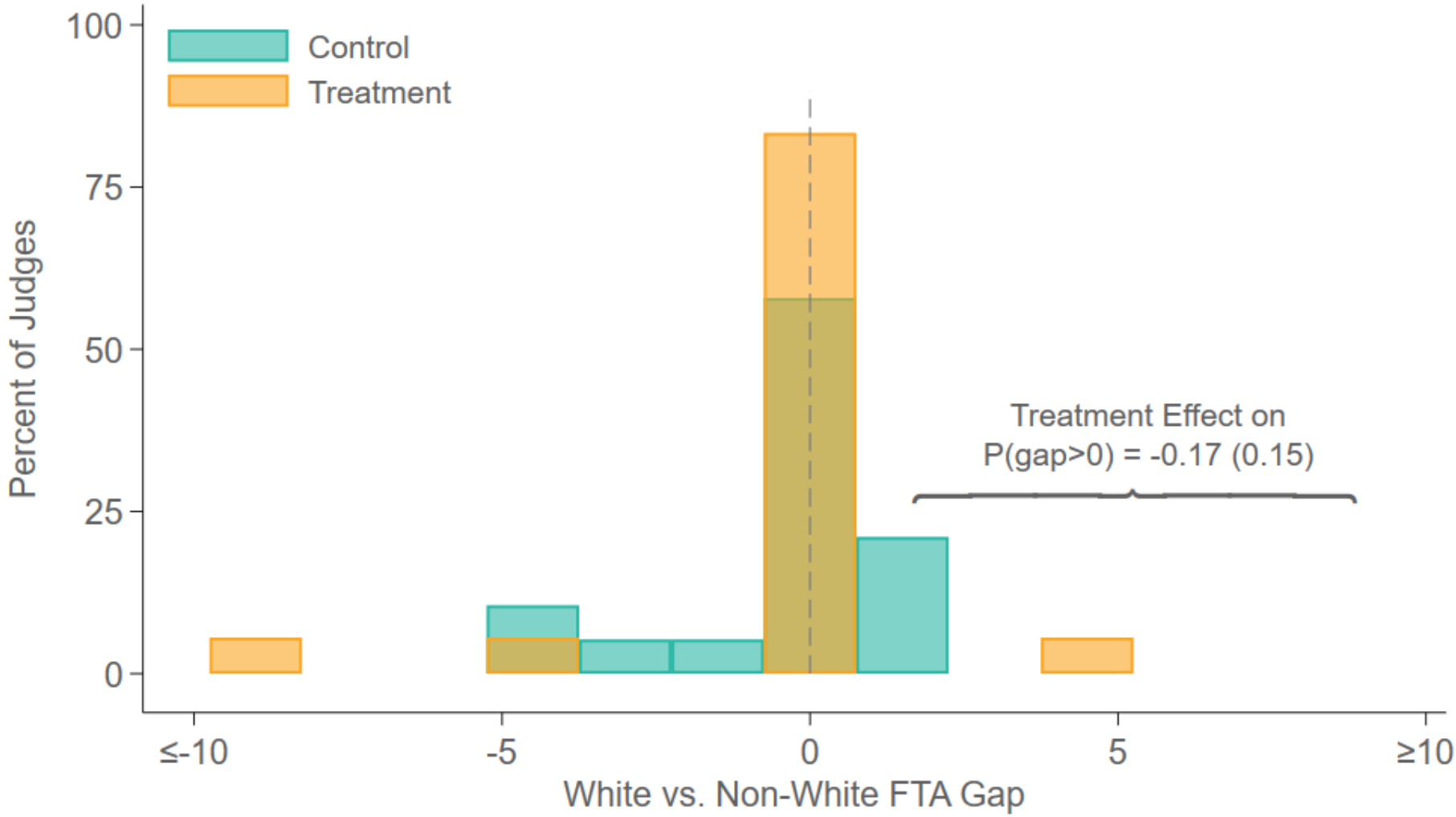
Note. This figure plots control vs. treatment judges' endline beliefs by race (N=37). Three judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects.

Estimation Error for Endline FTA Beliefs



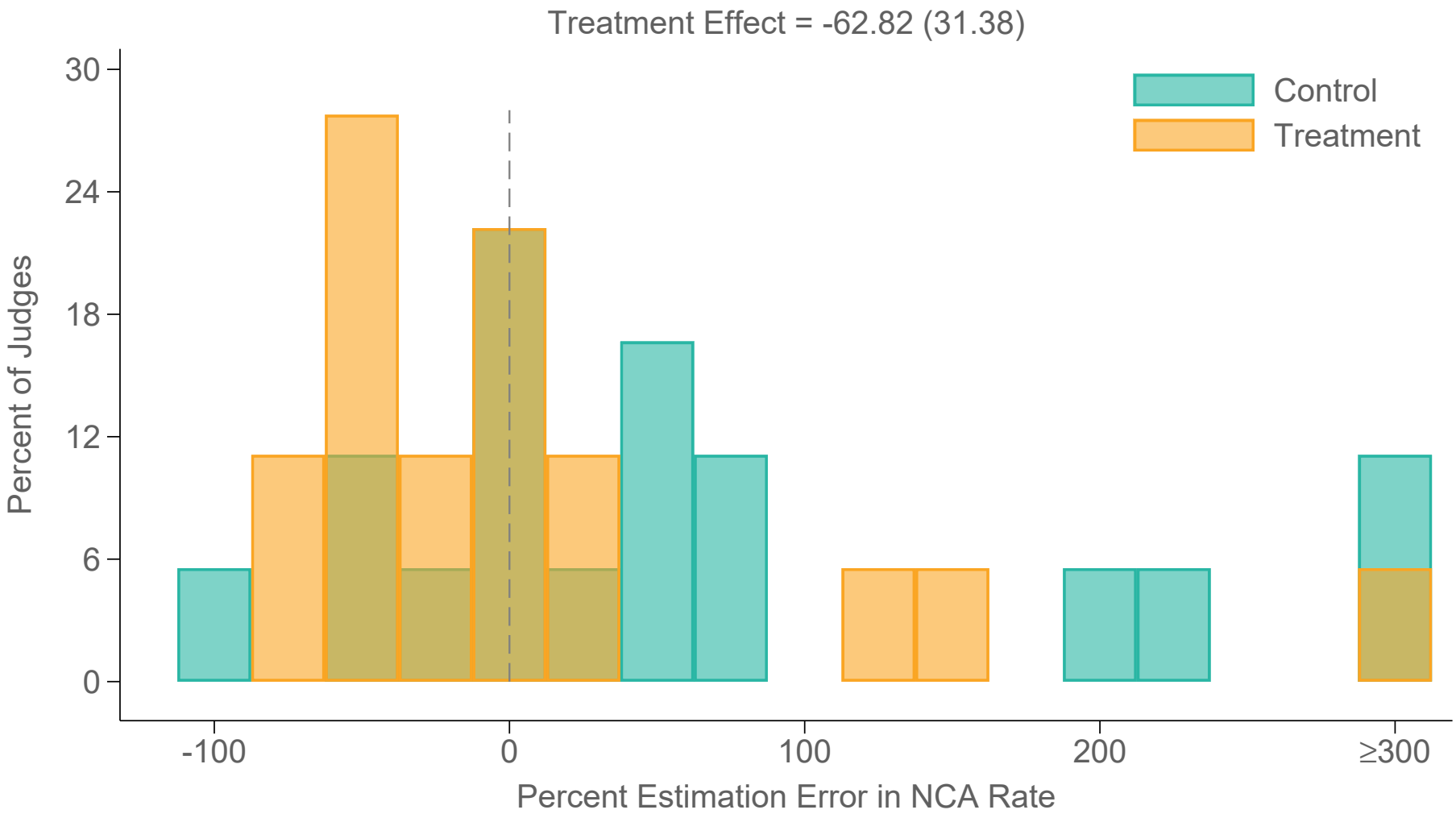
Note. This figure plots control vs. treatment judges' endline beliefs for all defendants (N=37). 3 judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects. ▸ White ▸ Non-White

Endline White vs. Non-White FTA Beliefs



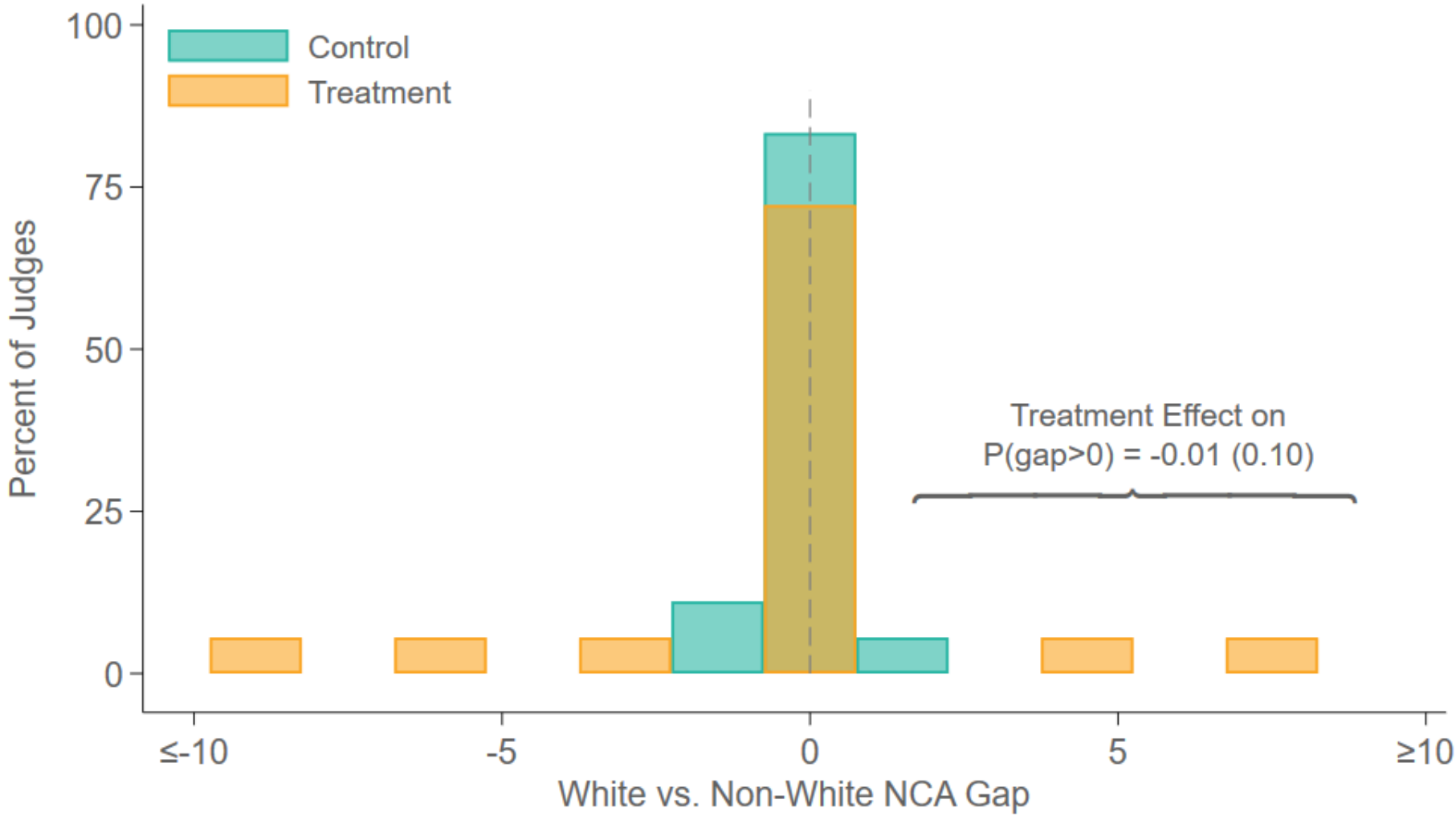
Note. This figure plots control vs. treatment judges' endline beliefs by race (N=37). Three judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects.

Estimation Error for Endline NCA Beliefs



Note. This figure plots control vs. treatment judges' endline beliefs for all defendants (N=36). 4 judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects. ▸ White ▸ Non-White

Endline White vs. Non-White NCA Beliefs

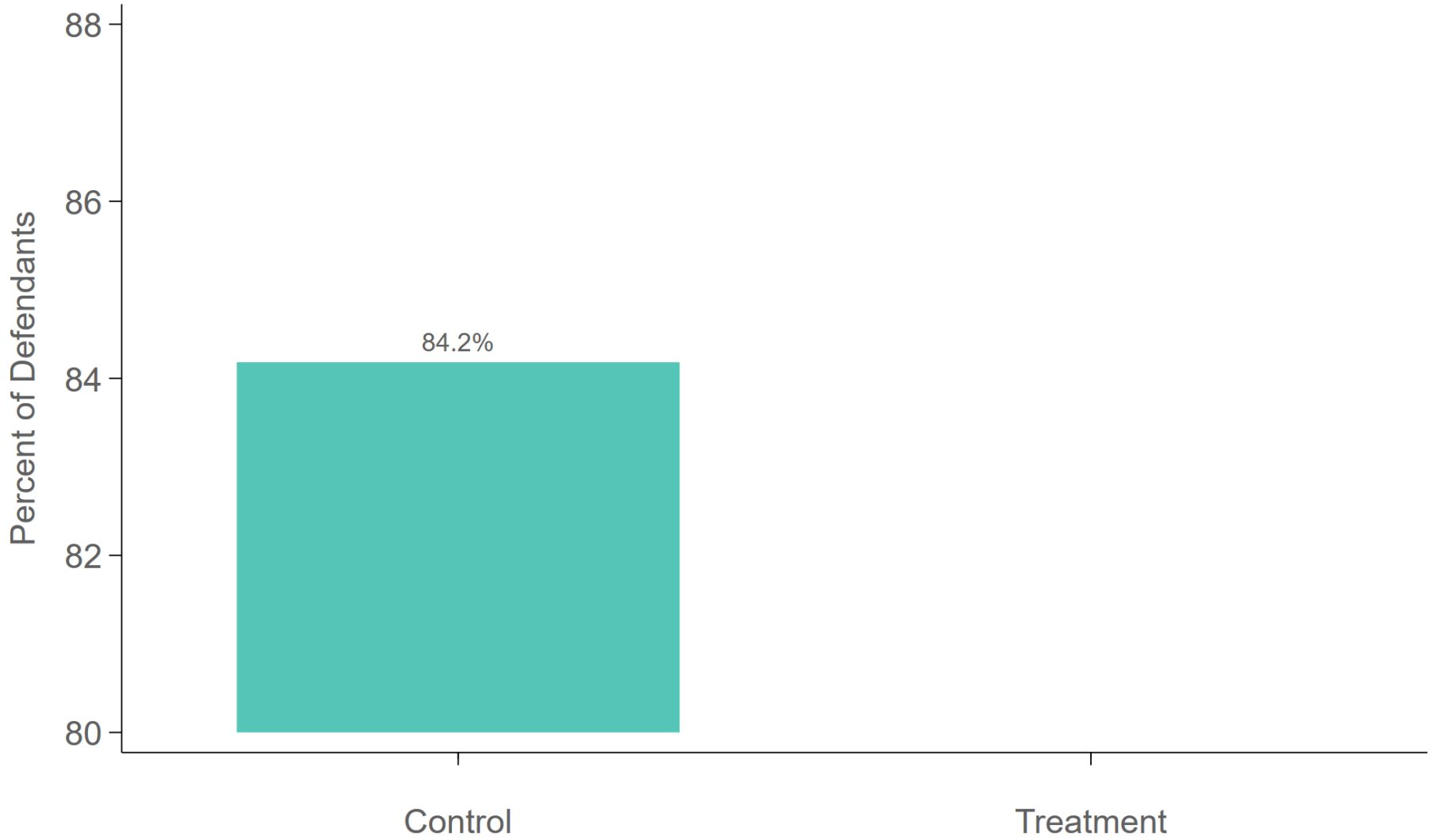


Note. This figure plots control vs. treatment judges' endline beliefs for all defendants (N=36). Four judges coded as "I don't know." Estimates come from a regression on treatment, baseline controls, and strata effects.

Roadmap

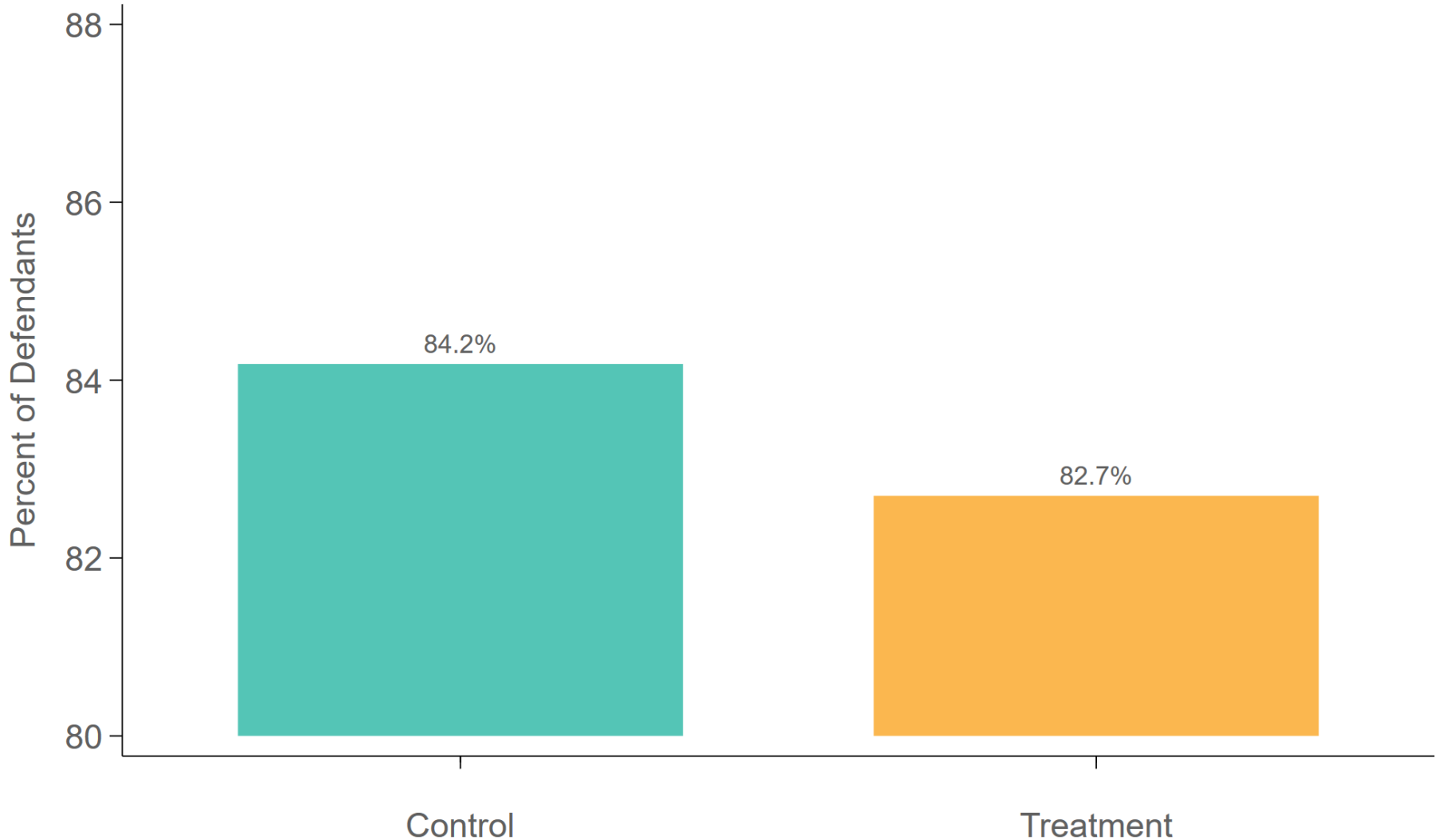
1. Taking Theory to the Field
2. Baseline Beliefs
3. Treatment Effects on Endline Beliefs
4. Treatment Effects on Pretrial Outcomes
5. Mechanisms
6. Conclusion

Release Rates



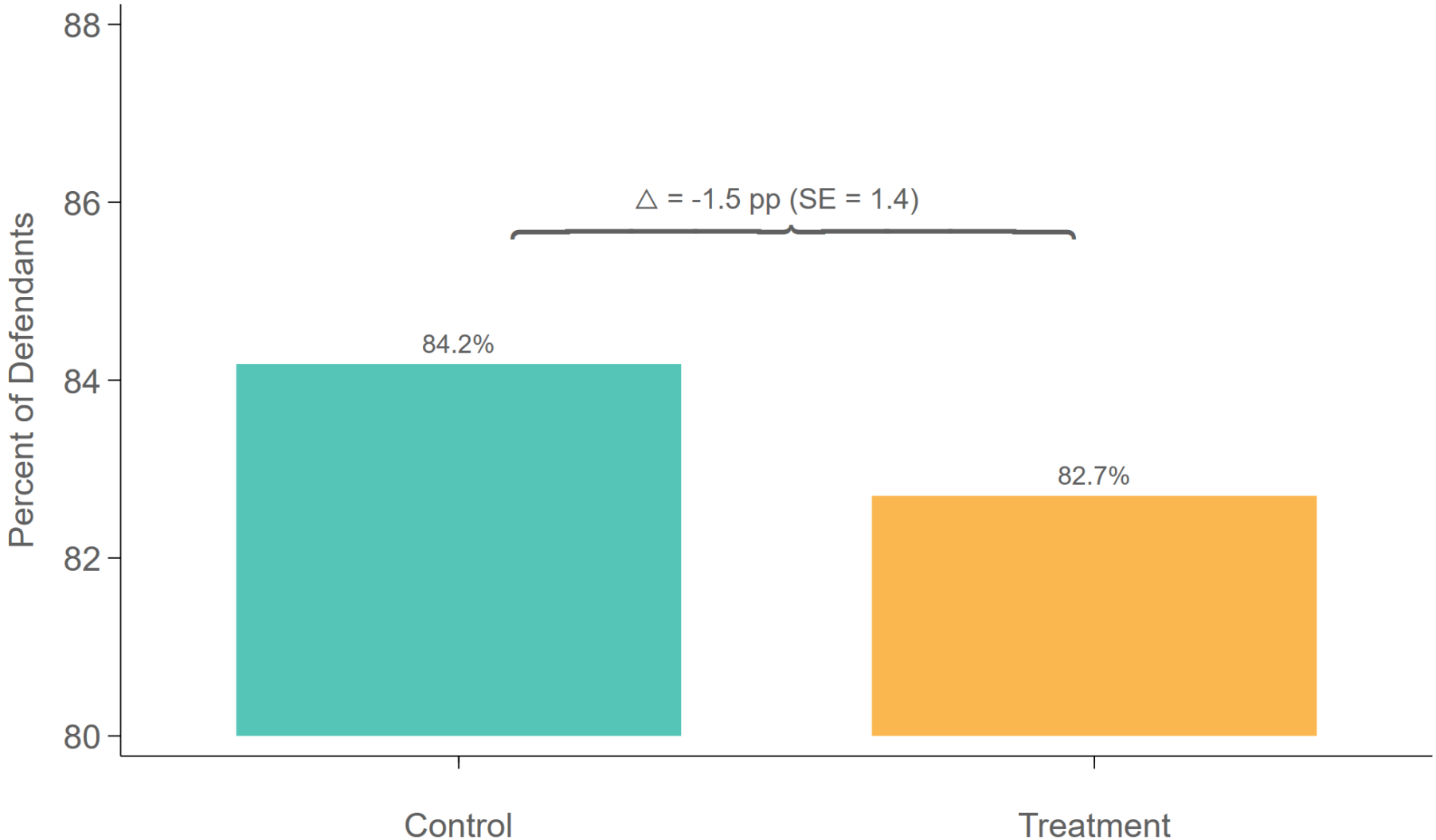
Note. This figure plots release outcomes by treatment status (N=51 judges, 30k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

Release Rates



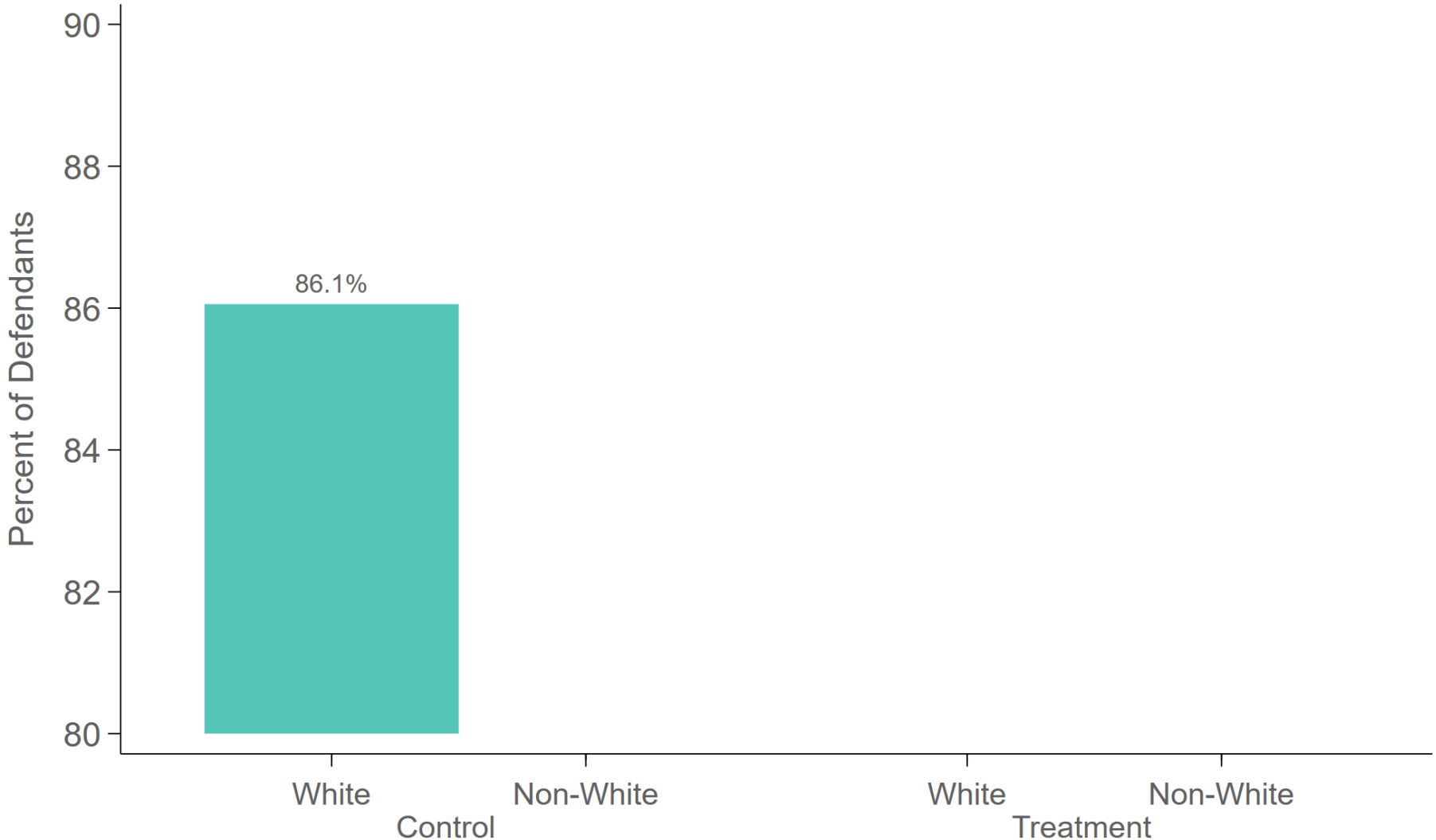
Note. This figure plots release outcomes by treatment status (N=51 judges, 30k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

Release Rates



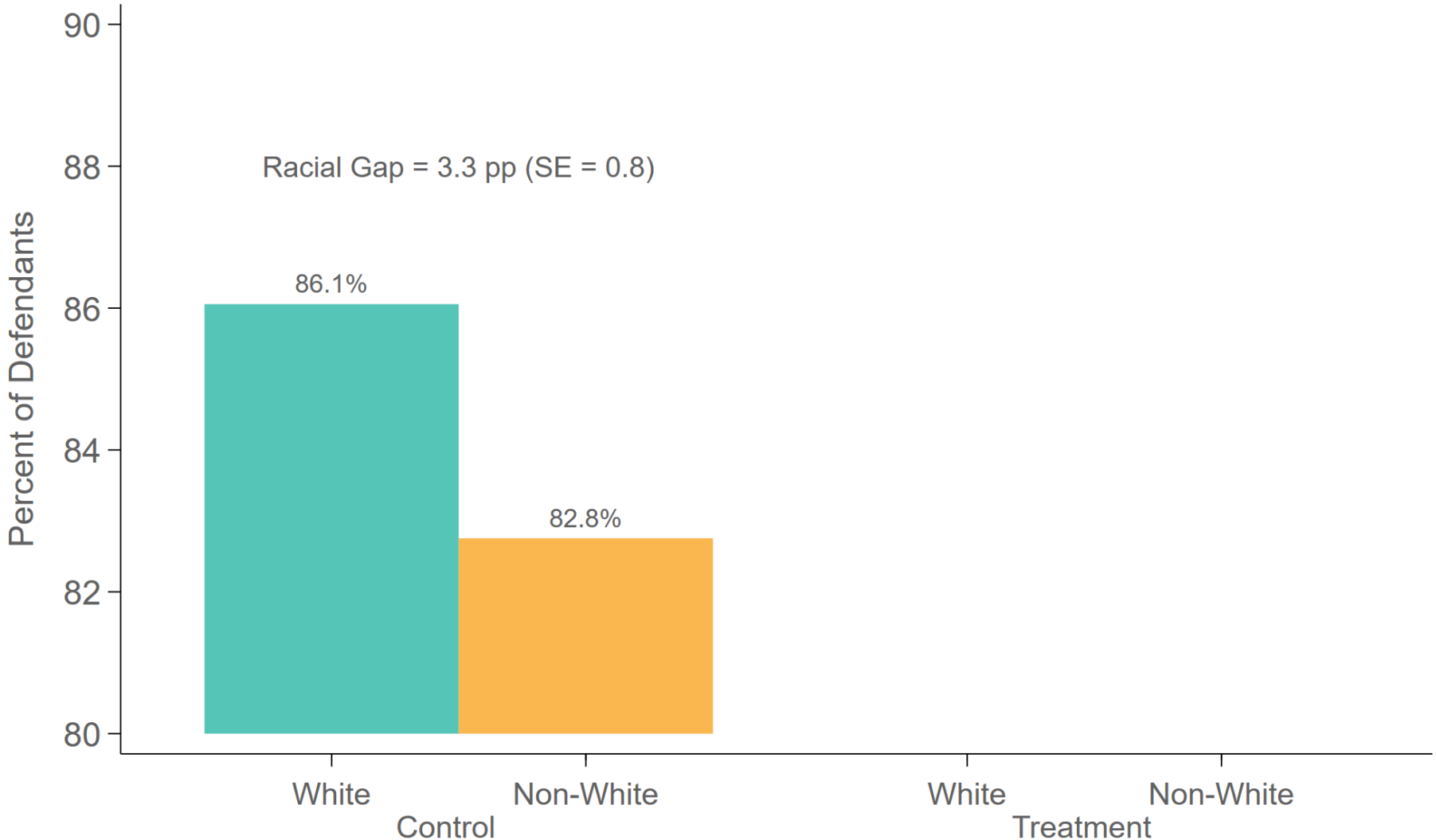
Note. This figure plots release outcomes by treatment status (N=51 judges, 30k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge. ▸ Misconduct

White vs. Non-White Release Rates



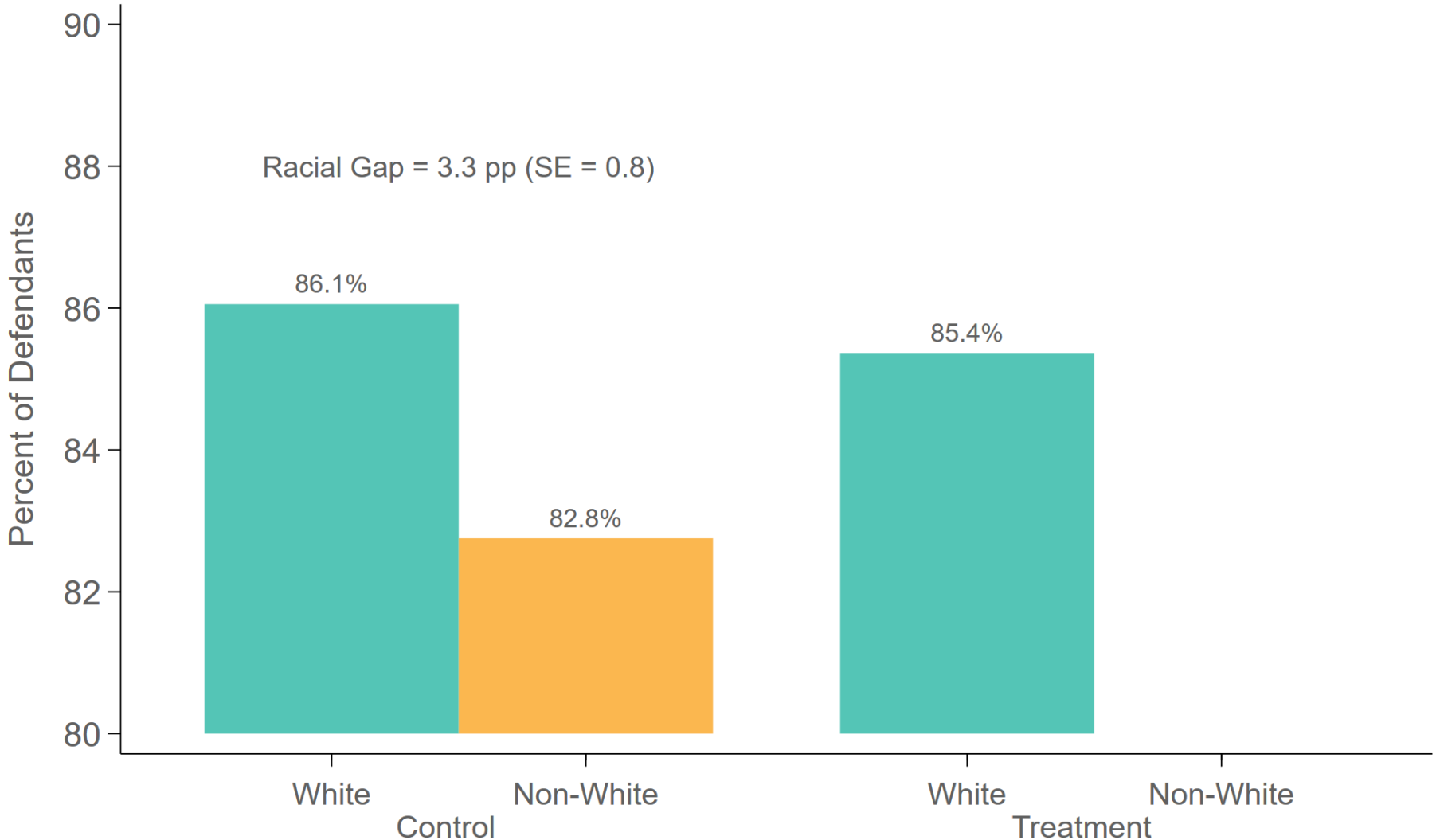
Note. This figure plots release outcomes by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Release Rates



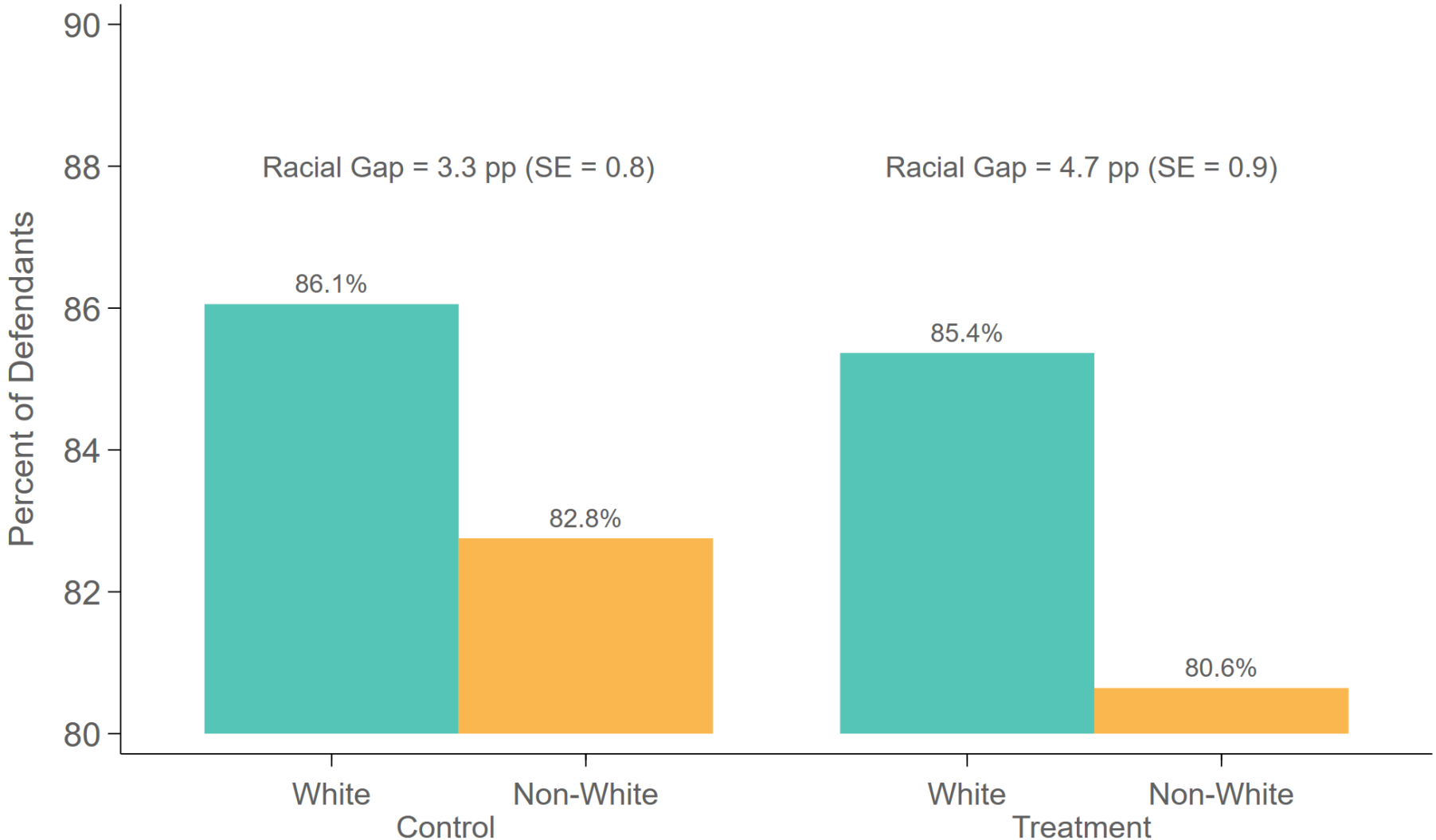
Note. This figure plots release outcomes by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Release Rates



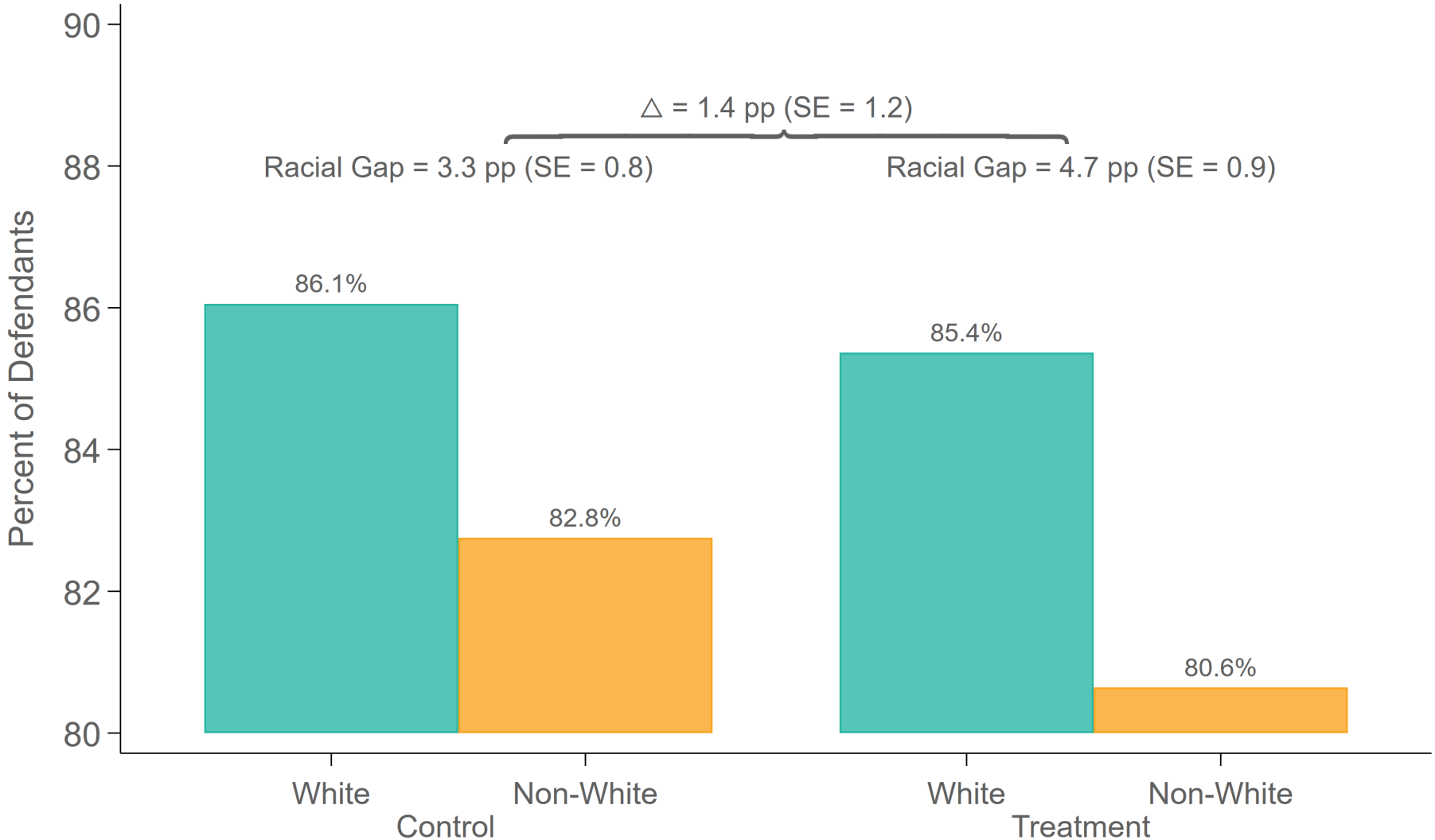
Note. This figure plots release outcomes by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Release Rates



Note. This figure plots release outcomes by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

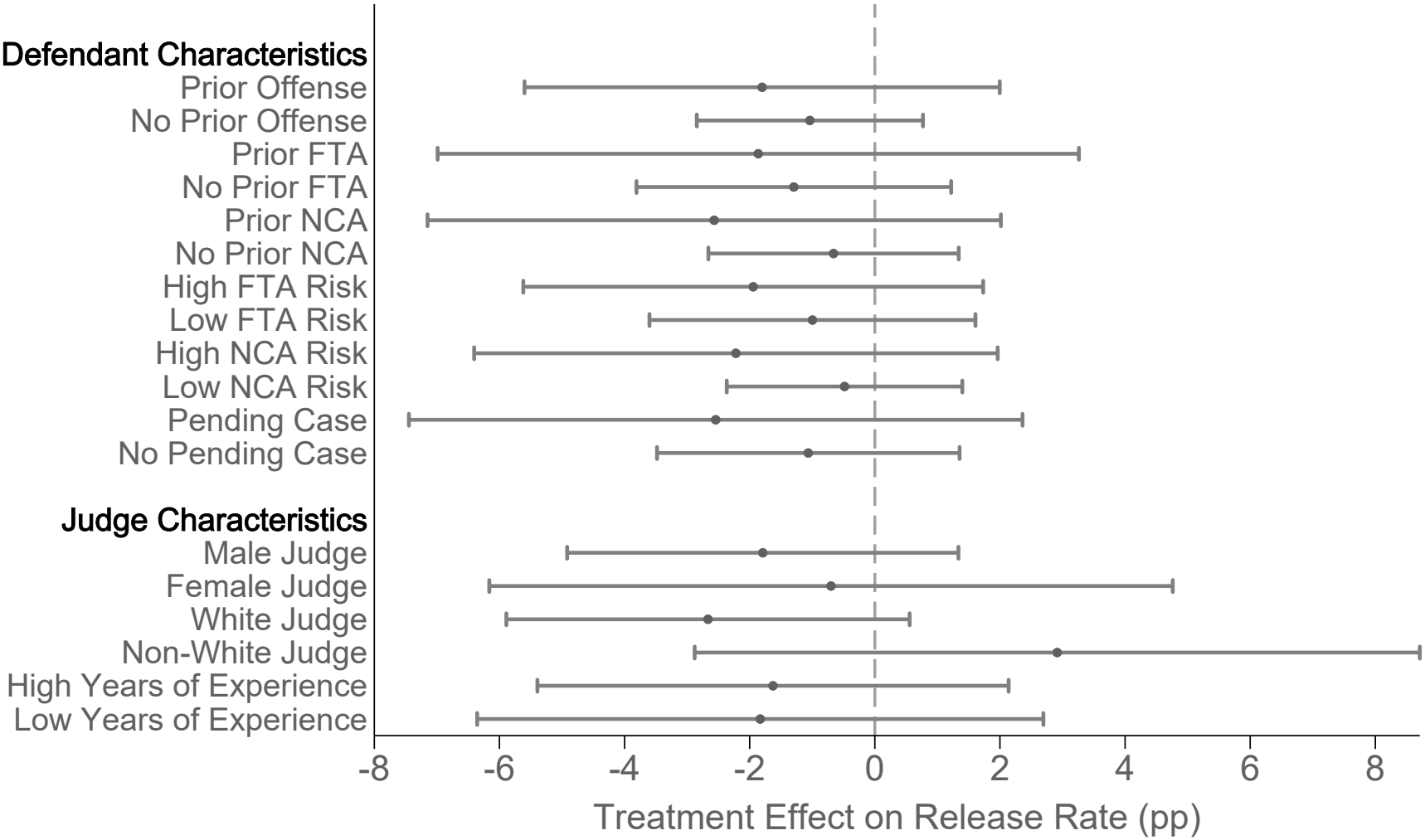
White vs. Non-White Release Rates



Note. This figure plots release outcomes by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

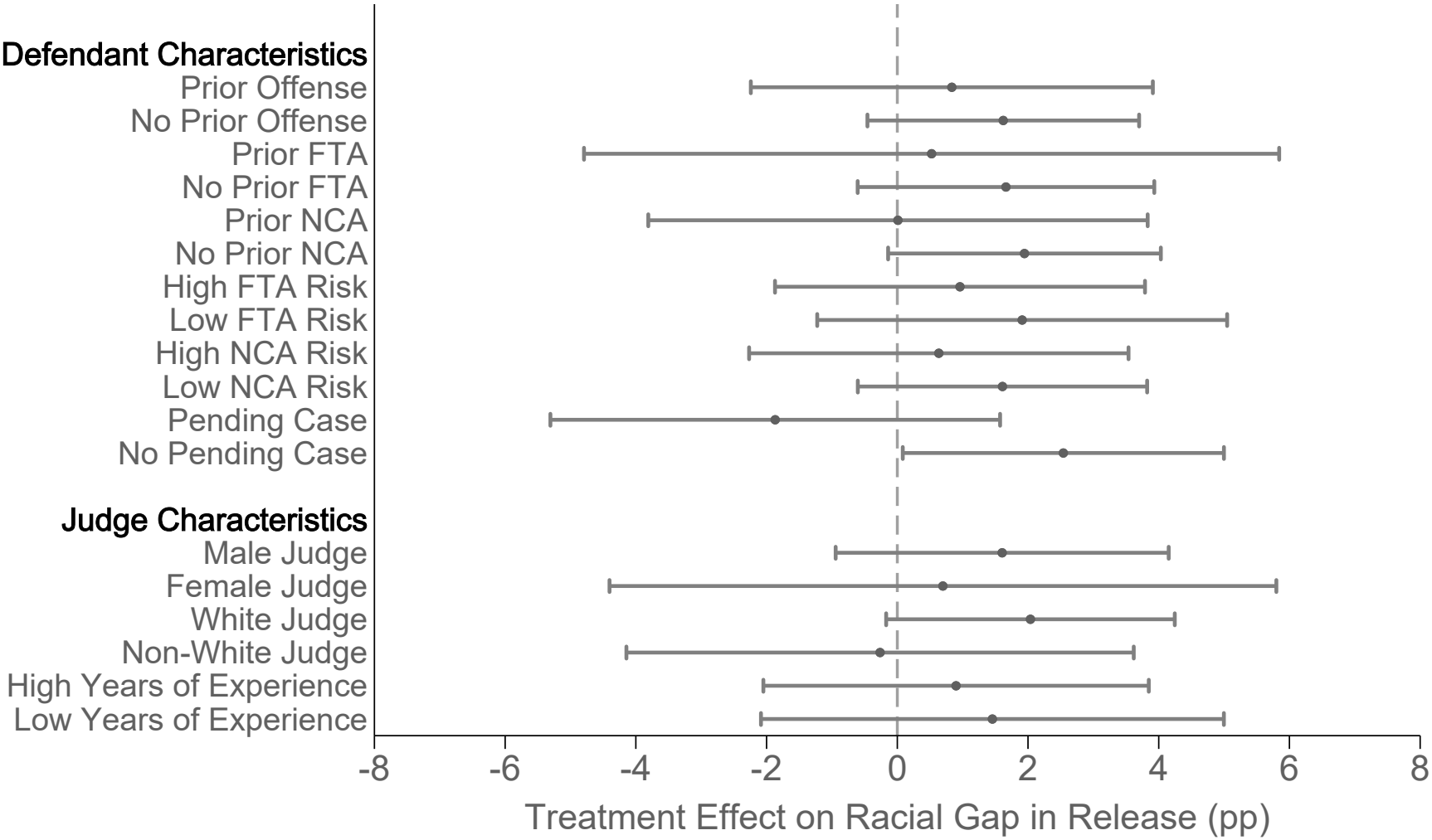
▸ Money Bail and ROR ▸ Misconduct

Release Subsample Results



Note. This figure reports the treatment effect on release rate by defendant and judge characteristics (N=51 judges, 30k cases). Point estimates and 95% confidence intervals are calculated from regressions of release on treatment status interacted with the indicated subgroup, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Release Subsample Results

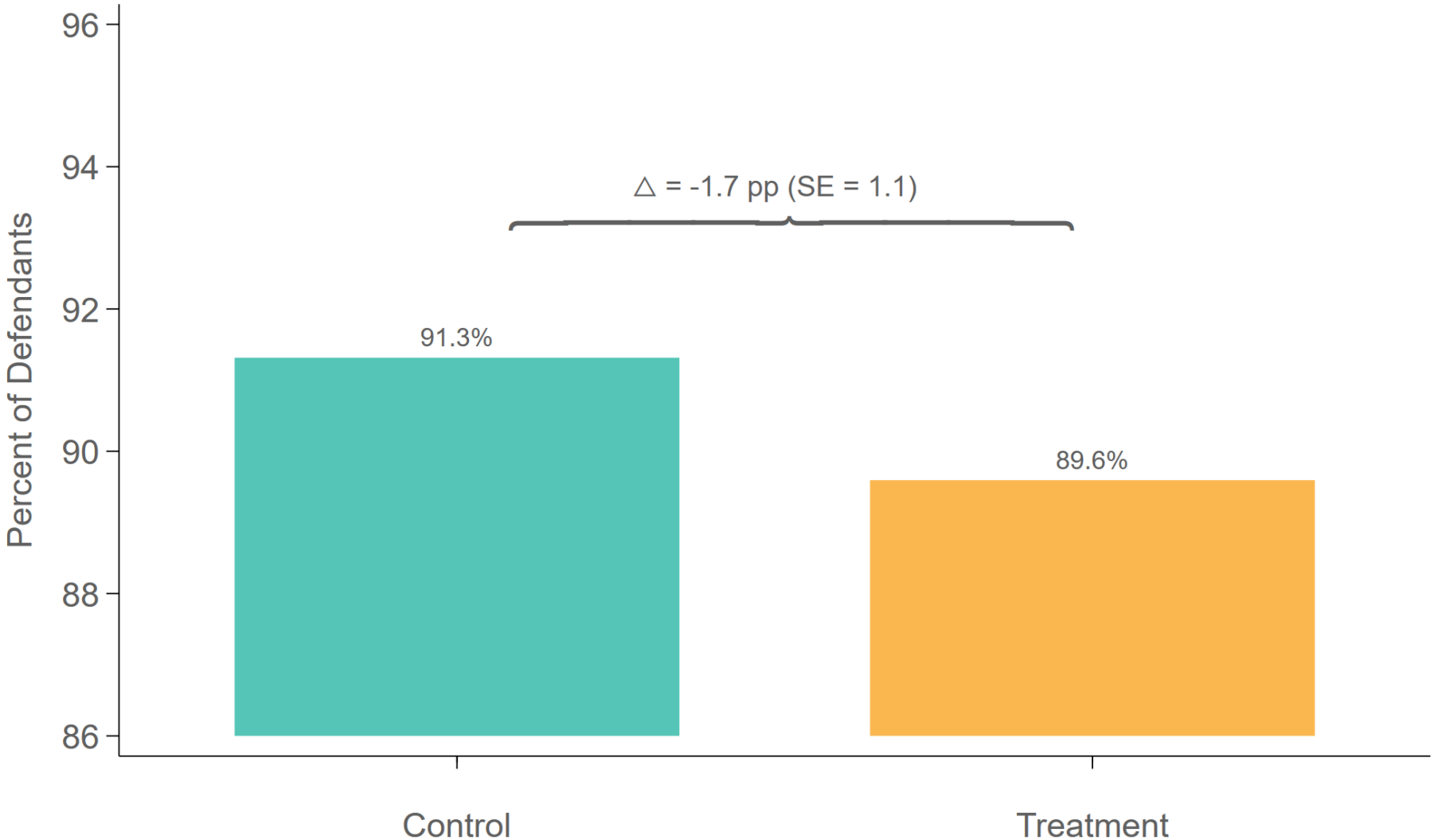


Note. This figure reports the treatment effect on the racial gap in release by defendant and judge characteristics (N=51 judges, 30k cases). Point estimates and 95% confidence intervals are calculated from regressions of release on treatment status interacted with race and the indicated subgroup, controlling for time effects. Standard errors are clustered by judge.

Follow-Up Experiment in New York City

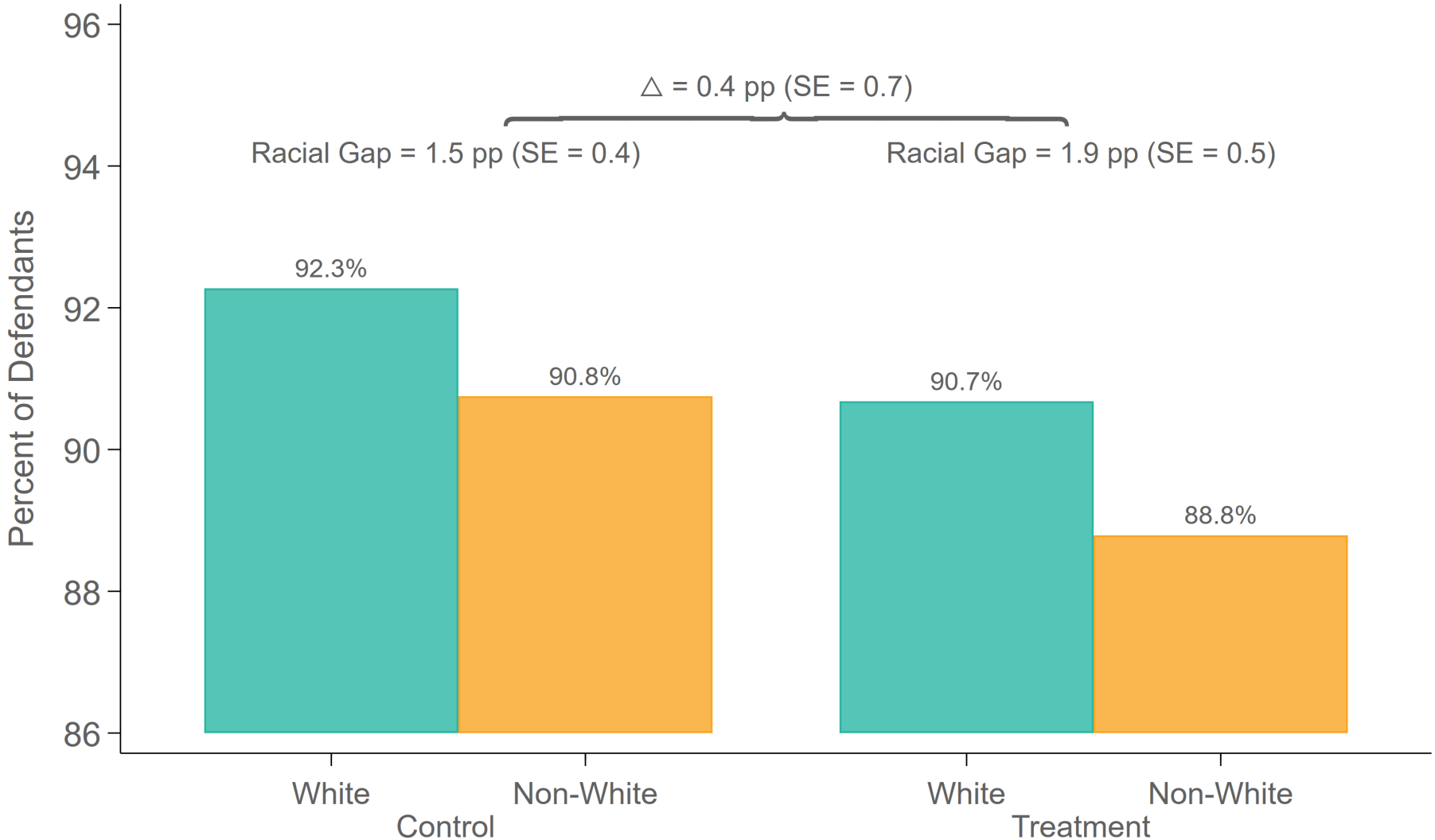
- We implemented a second experiment using publicly available data in NYC to explore external validity in a larger and non-selected sample
- Treated judges received one-time individualized feedback on their release and misconduct rates, racial gap in release rates, and a comparison to the average judge ▸ Intervention ▸ Summary Statistics
- Intervention materials sent in February 2024 to treated judges via email and FedEx, with required signature and notice of delivery
- We estimate treatment effects on release and misconduct using the 95 out of 106 judges with at least one case after the intervention

Release Rates in New York City



Note. This figure plots release outcomes by treatment status (N=95 judges, 56k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for time effects. Standard errors are clustered by judge. ▸ Misconduct

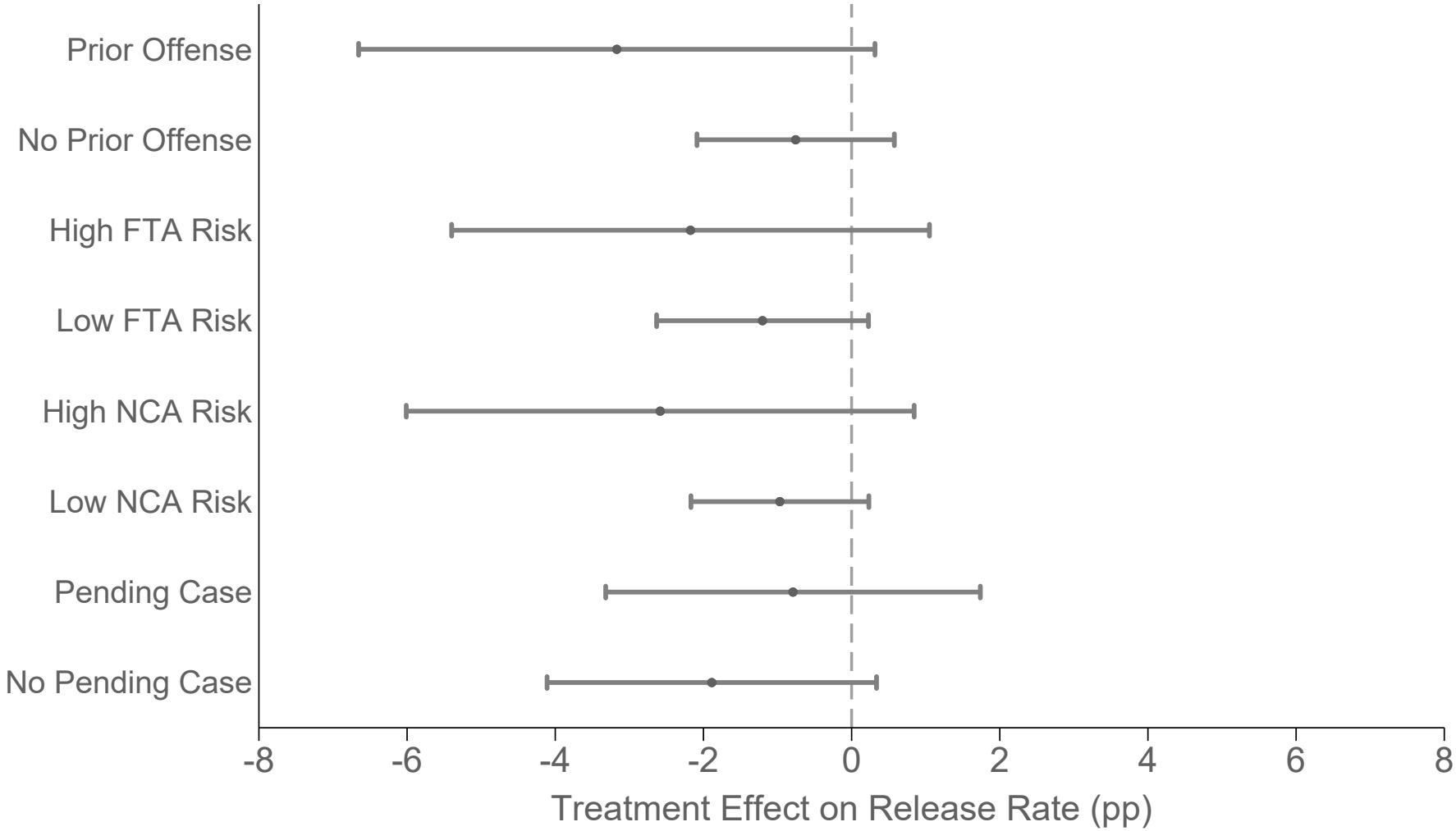
White vs. Non-White Release Rates in New York City



Note. This figure plots release outcomes by treatment status and defendant race (N=95 judges, 56k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for time effects. Standard errors are clustered by judge.

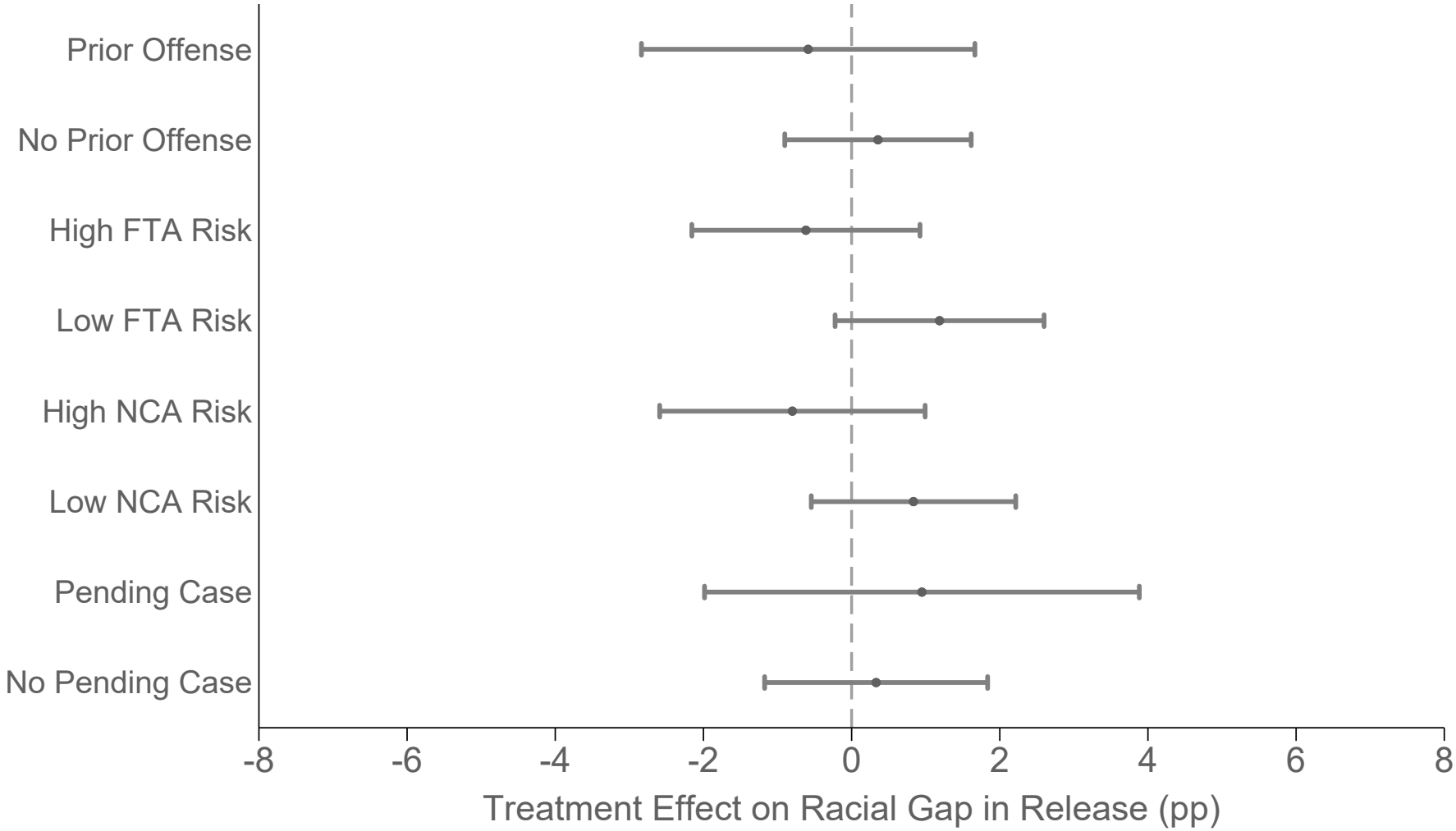
▸ Misconduct

Release Results in New York City



Note. This figure reports the treatment effect on release rate by defendant and judge characteristics (N=95 judges, 56k cases). Point estimates and 95% confidence intervals are calculated from regressions of release on treatment status interacted with the indicated subgroup, controlling for time effects. Standard errors are clustered by judge.

White vs. Non-White Release Results in New York City



Note. This figure reports the treatment effect on the racial gap in release by defendant and judge characteristics (N=95 judges, 56k cases). Point estimates and 95% confidence intervals are calculated from regressions of release on treatment status interacted with race and the indicated subgroup, controlling for time effects. Standard errors are clustered by judge.

Roadmap

1. Taking Theory to the Field
2. Baseline Beliefs
3. Treatment Effects on Endline Beliefs
4. Treatment Effects on Pretrial Outcomes
5. Mechanisms
6. Conclusion

Taking Stock

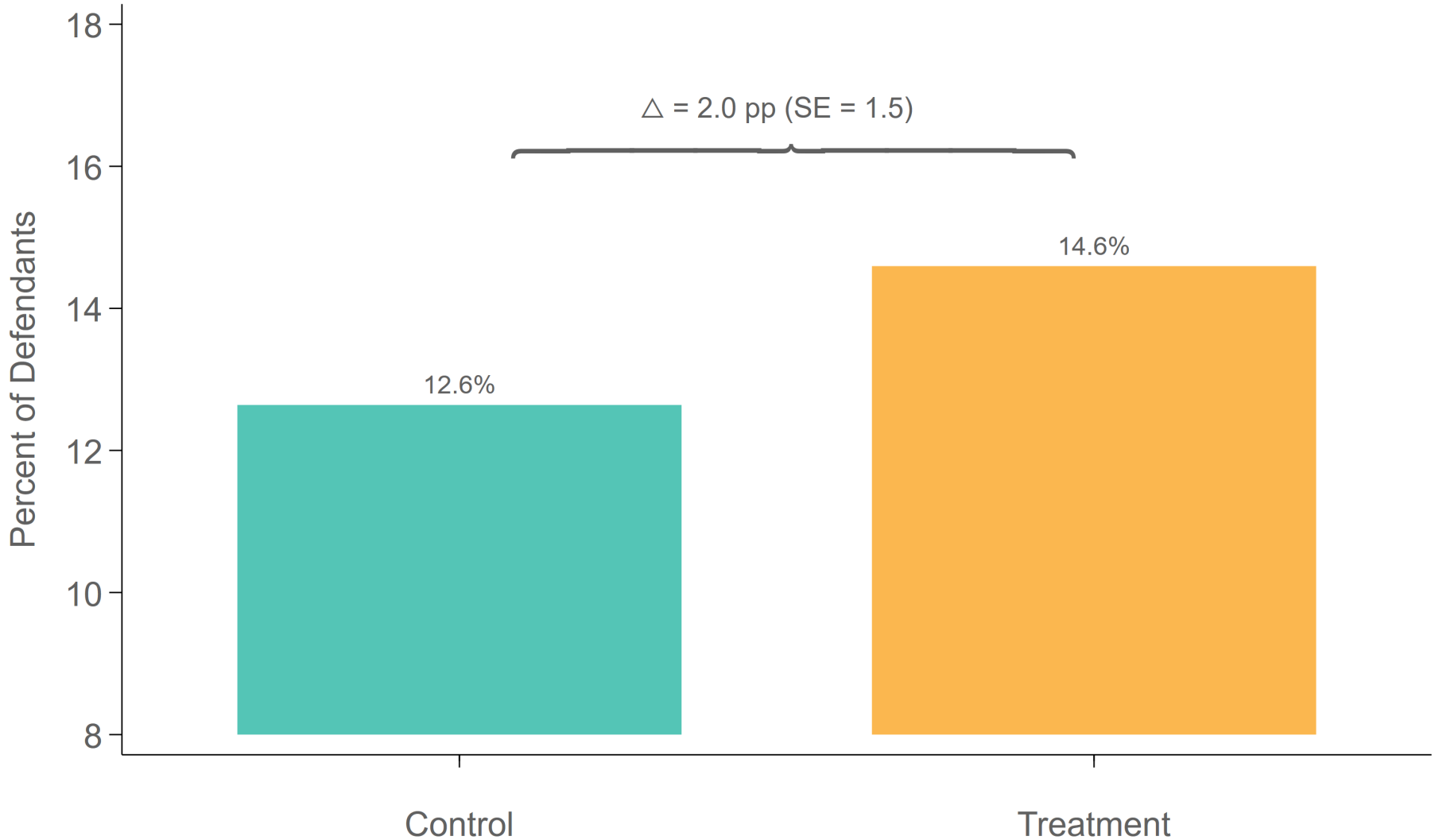
- Treated judges update their beliefs, with particularly large reductions in the perceived risk of releasing Non-White defendants
- Standard models generally predict that these changes should lead to *higher* release rates and *lower* White vs. Non-White release disparities
- Yet we find *no change* in release rates and White vs. Non-White release disparities, rejecting the predictions from standard models

Potential Explanations

We explore three potential explanations for our results:

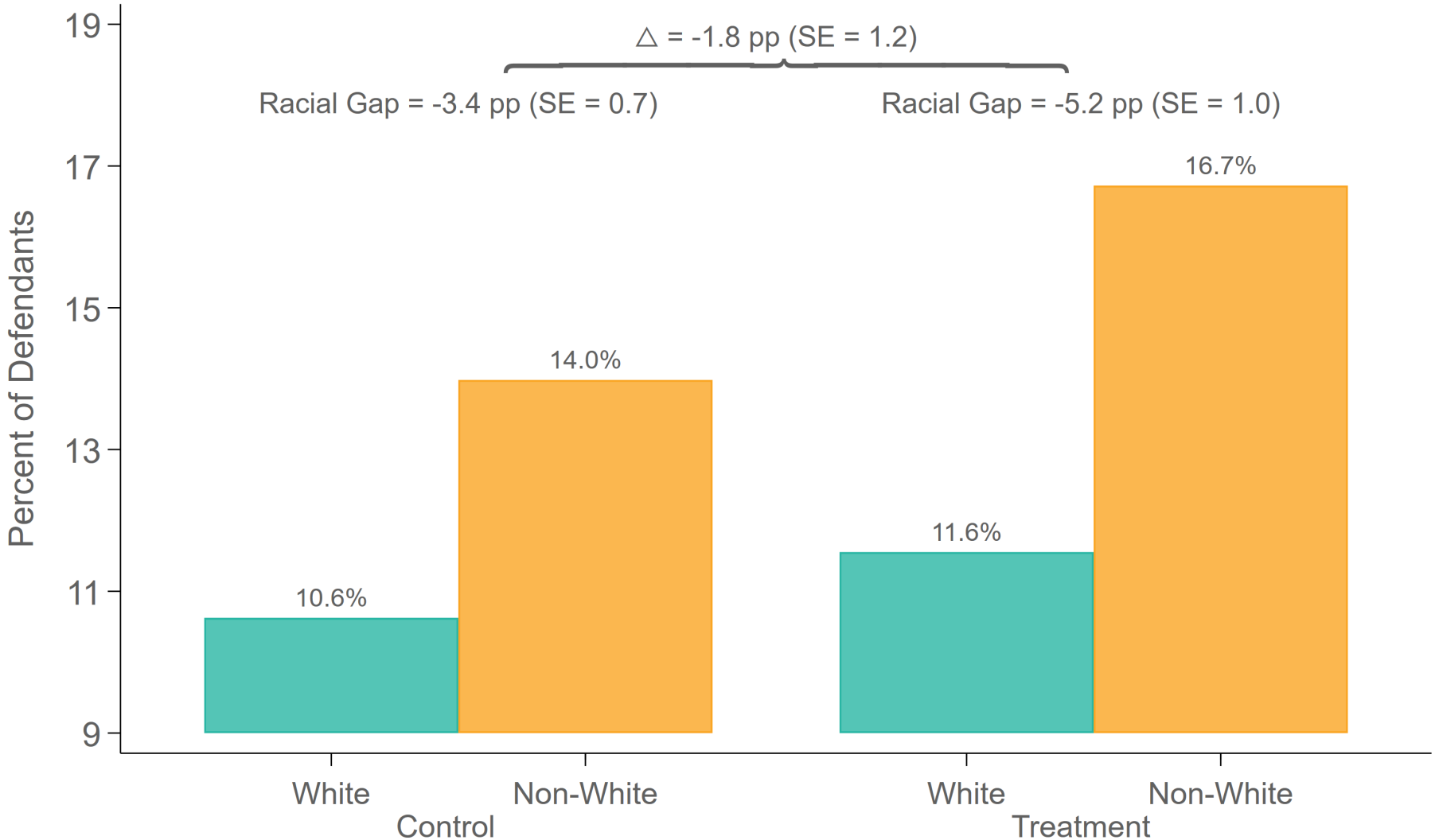
1. Judges need specific guidance on each case to change their release decisions, not just court statistics and individualized feedback
2. Increasing the salience of racial disparities and/or misconduct may lead to backlash or other unintended consequences
3. Strong preferences for releasing defendants (i.e., low Type II/Type I error costs) may result in some judges not changing their release rates

Algorithmic Overrides to Detain Low-Risk Defendants



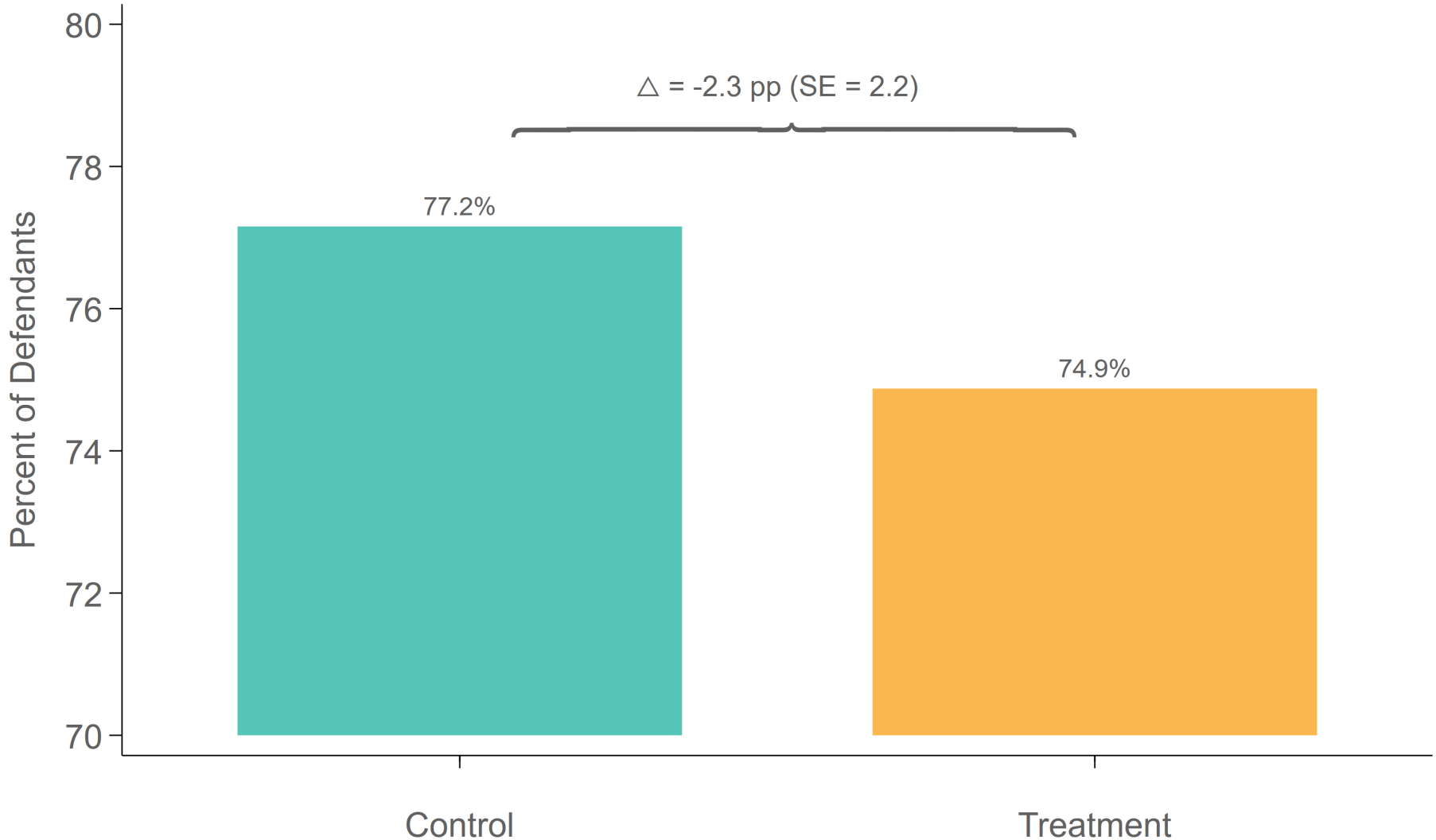
Note. This figure plots judge overrides to detain when the algorithm recommends release (N=37 judges, 21k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Overrides to Detain Low-Risk



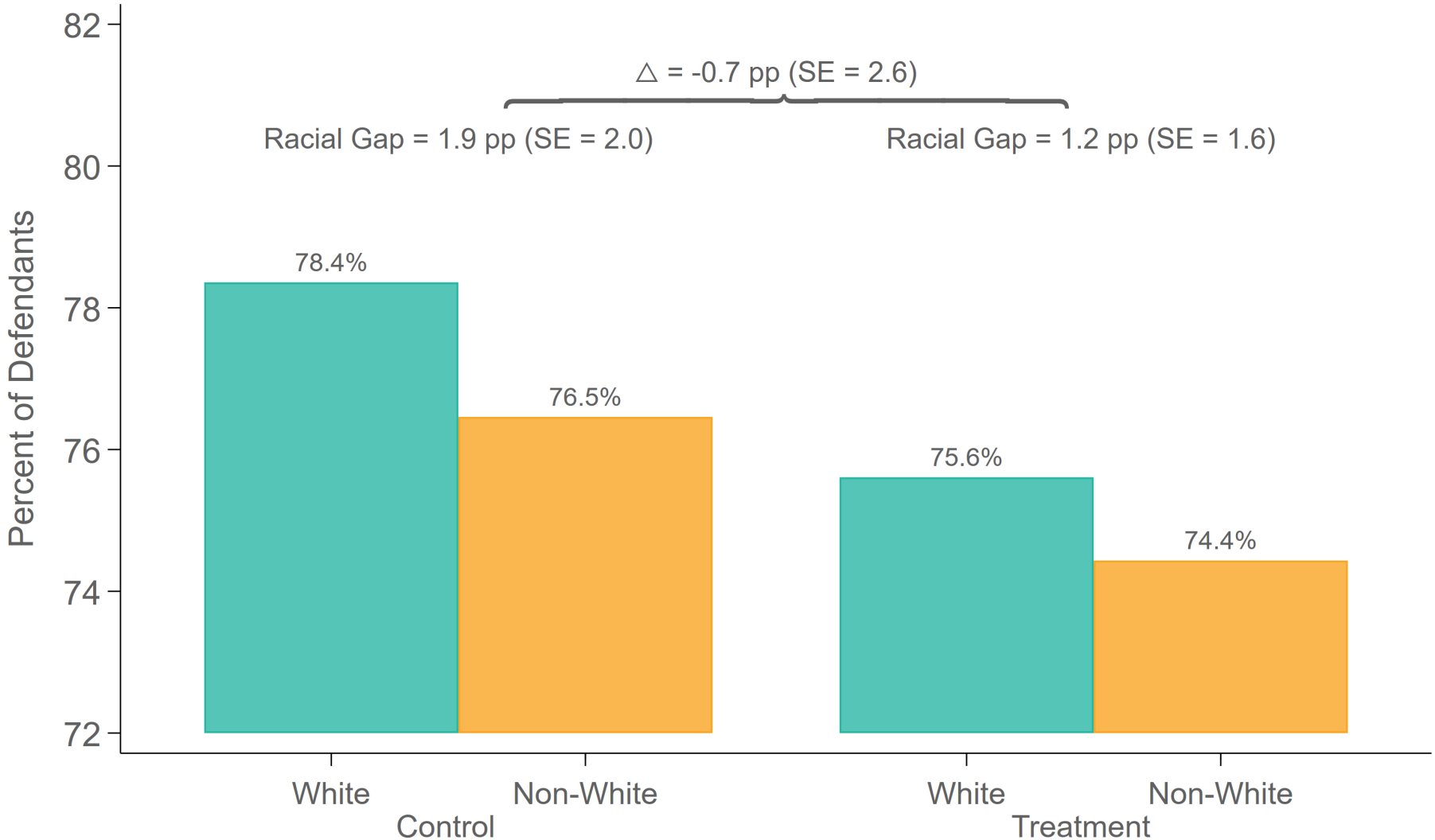
Note. This figure plots judge overrides to detain when the algorithm recommends release (N=37 judges, 21k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for time effects. Standard errors are clustered by judge.

Algorithmic Overrides to Release High-Risk Defendants



Note. This figure plots judge overrides to release when the algorithm recommends detention (N=37 judges, 4.2k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Overrides to Release High-Risk



Note. This figure plots judges overrides to release when the algorithm recommends detention (N=37 judges, 4.2k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for time effects. Standard errors are clustered by judge.

Potential Explanations

We explore three potential explanations for our results:

X 1. Judges need specific guidance on each case to change their release decisions, not just court statistics and individualized feedback

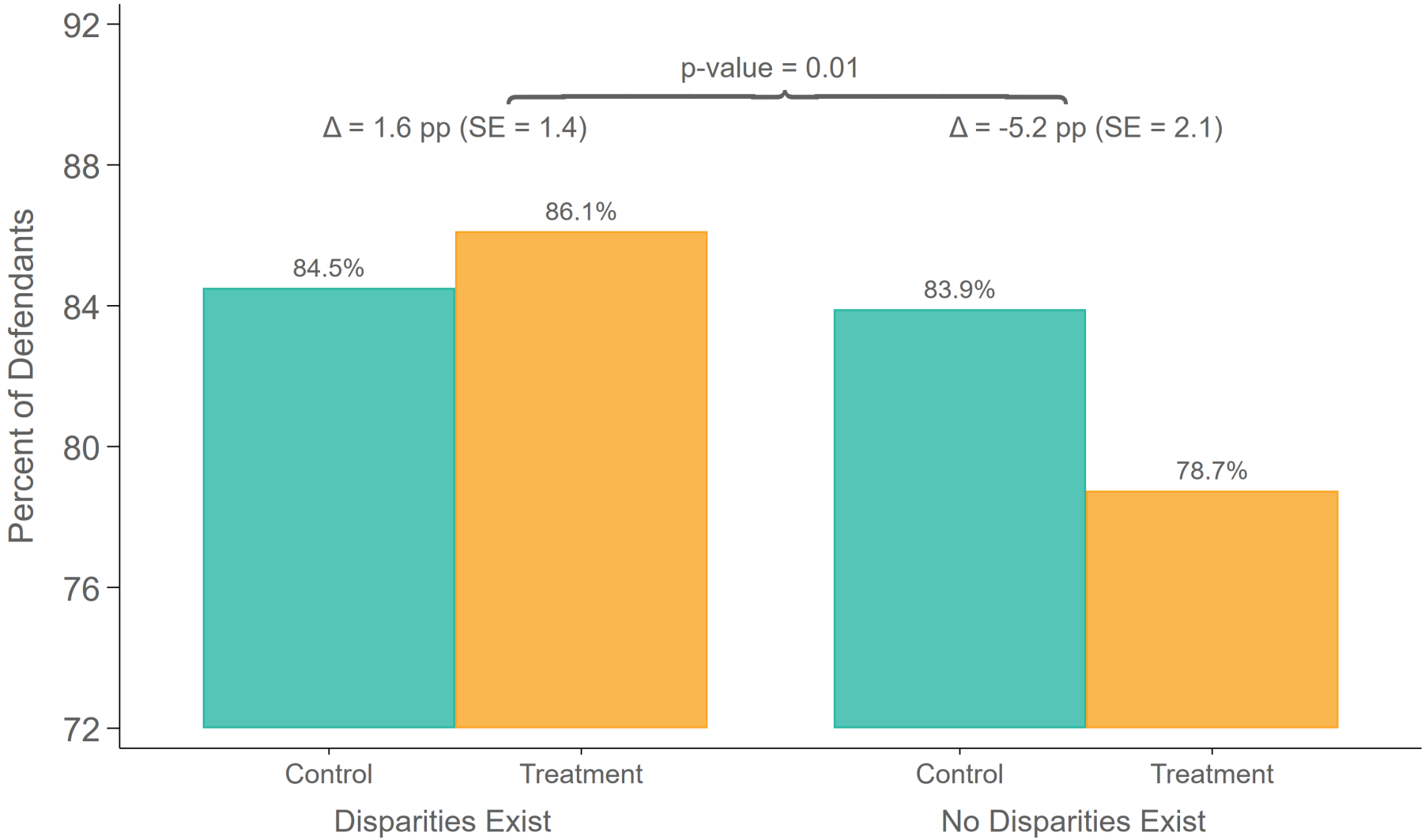
2. Increasing the salience of racial disparities and/or misconduct may lead to backlash or other unintended consequences

3. Strong preferences for releasing defendants (i.e., low Type II/Type I error costs) may result in some judges not changing their release rates

Behaviorally-Based Frictions

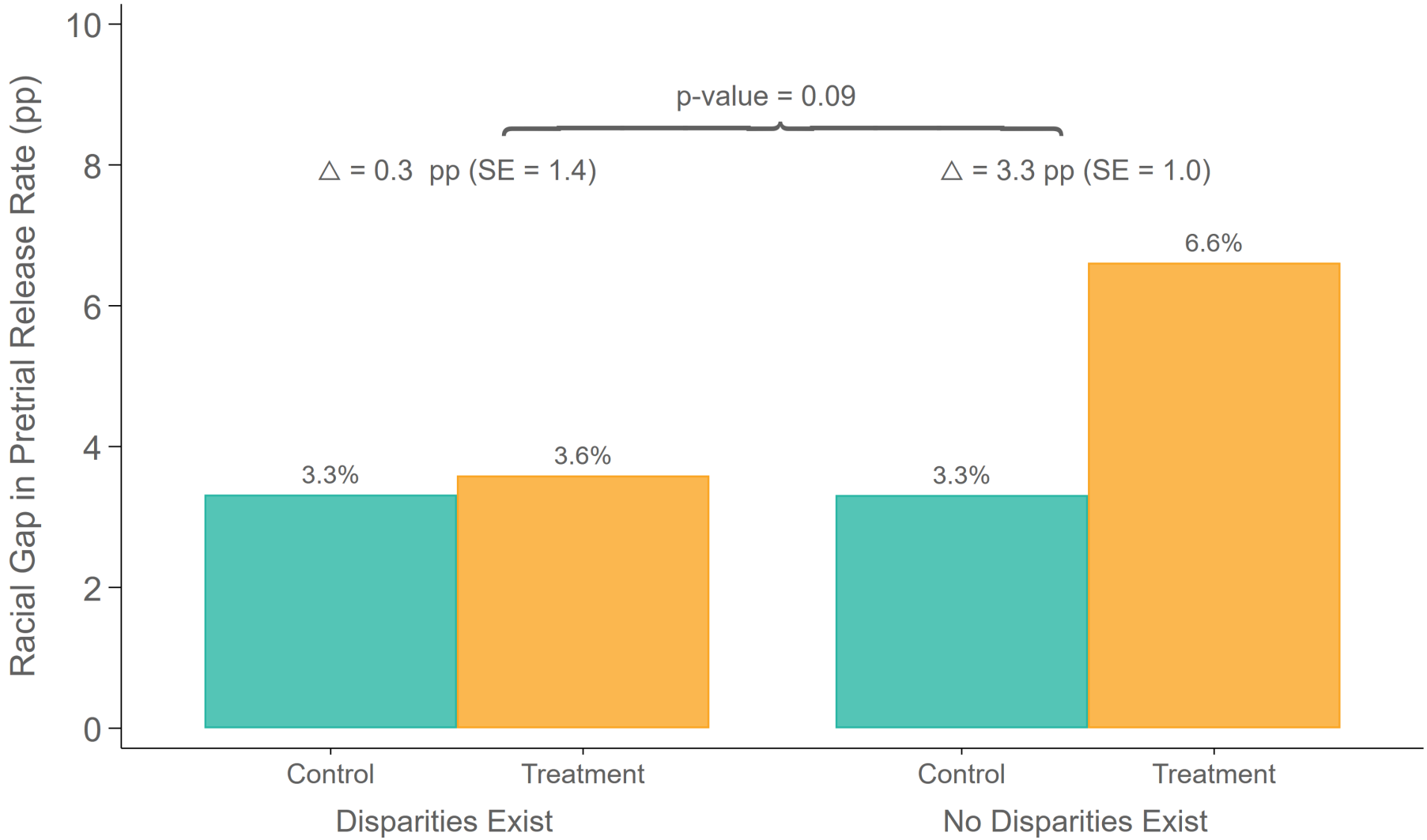
- Prior work shows that highlighting group differences can increase prejudice and reactance (Paluck 2006, Paluck and Green 2009)
- We use two questions to proxy for a greater potential for backlash and reactance among the judges in our sample:
  1. Believes that there are no racial disparities among their peers
  2. Believes that there are no racial disparities in pretrial generally
- Both questions are positively correlated with anti-Black IAT scores and proxy for views that may be challenged by our intervention

Release Rates by Baseline Disparity Beliefs



Note. This figure plots release by treatment status and baseline beliefs about release disparities in peers' decisions and in the pretrial system. We report the control mean for judges reporting at least one disparity; other estimates come from a regression of case-level release on treatment interacted with an indicator for reporting no disparities, controlling for court-by-time effects. Standard errors are clustered by judge. ▸ Baseline IAT

Release Rate Gaps by Baseline Disparity Beliefs



Note. This figure plots the racial gap in release by treatment status and baseline beliefs about release disparities in peers' decisions and in the pretrial system. We report the control mean for judges reporting at least one disparity; other estimates come from a regression of case-level release on treatment interacted with race and an indicator for reporting no disparities, controlling for court-by-time effects. Standard errors are clustered by judge. ▸ Baseline IAT

Potential Explanations

We explore three potential explanations for our results:

X 1. Judges need specific guidance on each case to change their release decisions, not just court statistics and individualized feedback

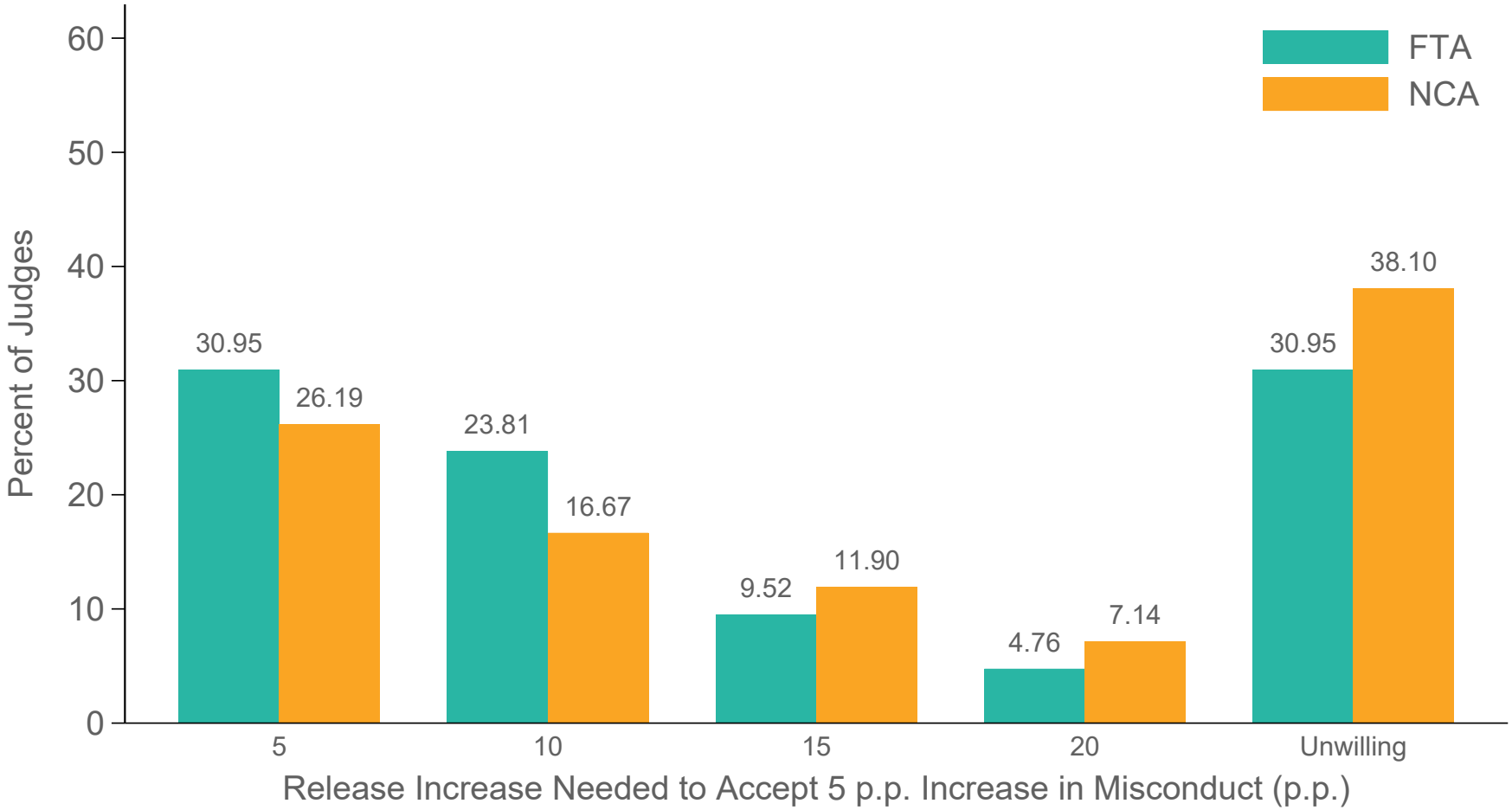
✓ 2. Increasing the salience of racial disparities and/or misconduct may lead to backlash or other unintended consequences

3. Strong preferences for releasing defendants (i.e., low Type II/Type I error costs) may result in some judges not changing their release rates

Preference-Based Frictions

- Standard models of decision-making generally suggest that correcting exaggerated beliefs about misconduct risk will increase release rates
- But preferences over Type II vs. Type I error costs may limit the impact of correcting these exaggerated beliefs on release rates
  - Judges with high Type II/Type I error costs (i.e., a "high" price for additional misconduct) will increase their release rates
  - But judges with low Type II/Type I error costs (i.e., a "low" price for additional misconduct) may not change their release rates
- We use baseline survey questions eliciting preferences to categorize judges into those with high vs. low Type II/Type I error costs

Baseline Preferences Over Type II vs. Type I Error Costs



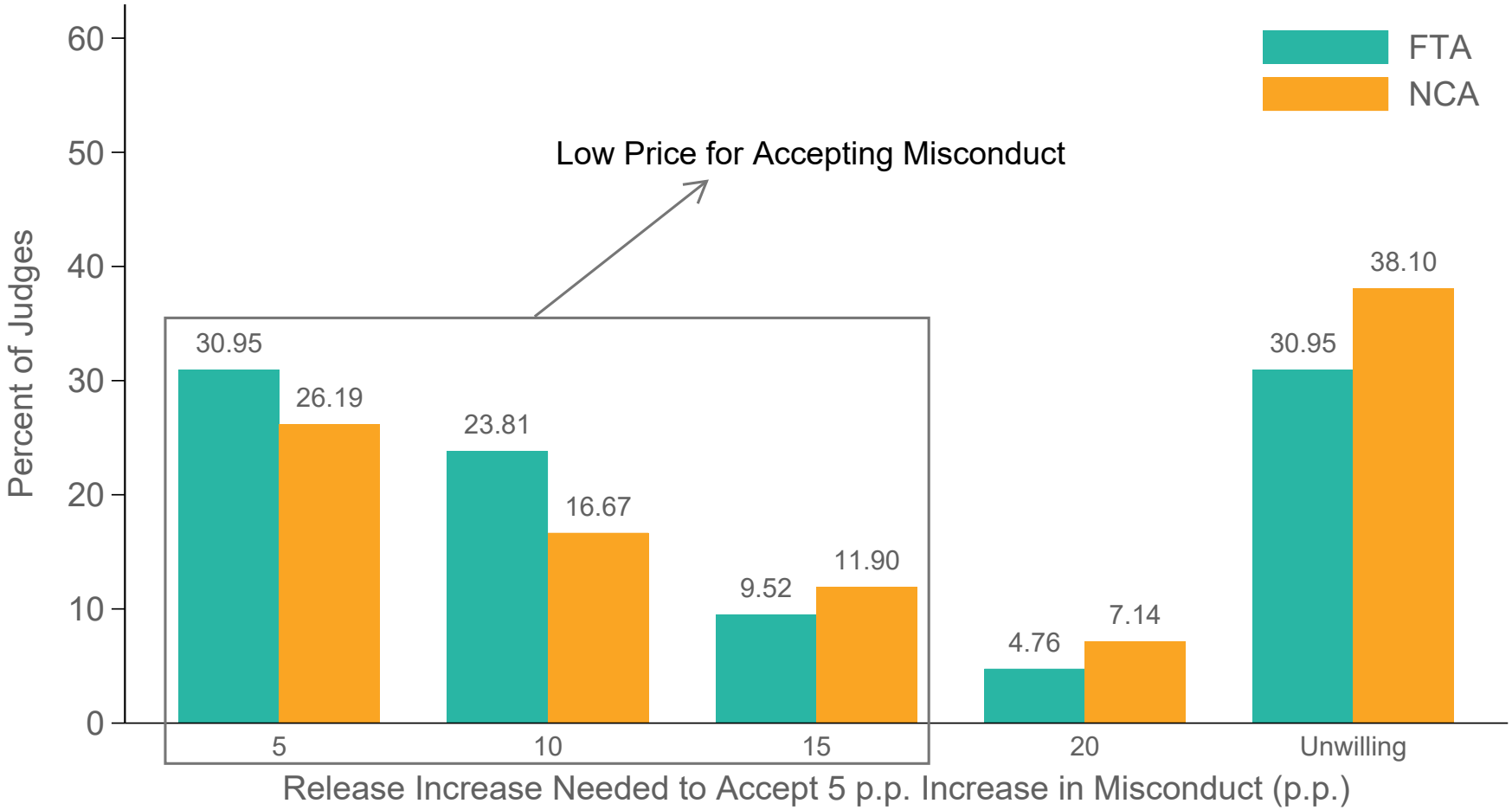
Note. This figure plots the distribution of judges by the release rate increase they require to accept a 5 percentage point increase in misconduct, relative to a baseline of 80 percent release and 20 percent misconduct. The unwilling category includes judges who never accept the higher misconduct rate even at 100 percent release.

Baseline Preferences Over Type II vs. Type I Error Costs



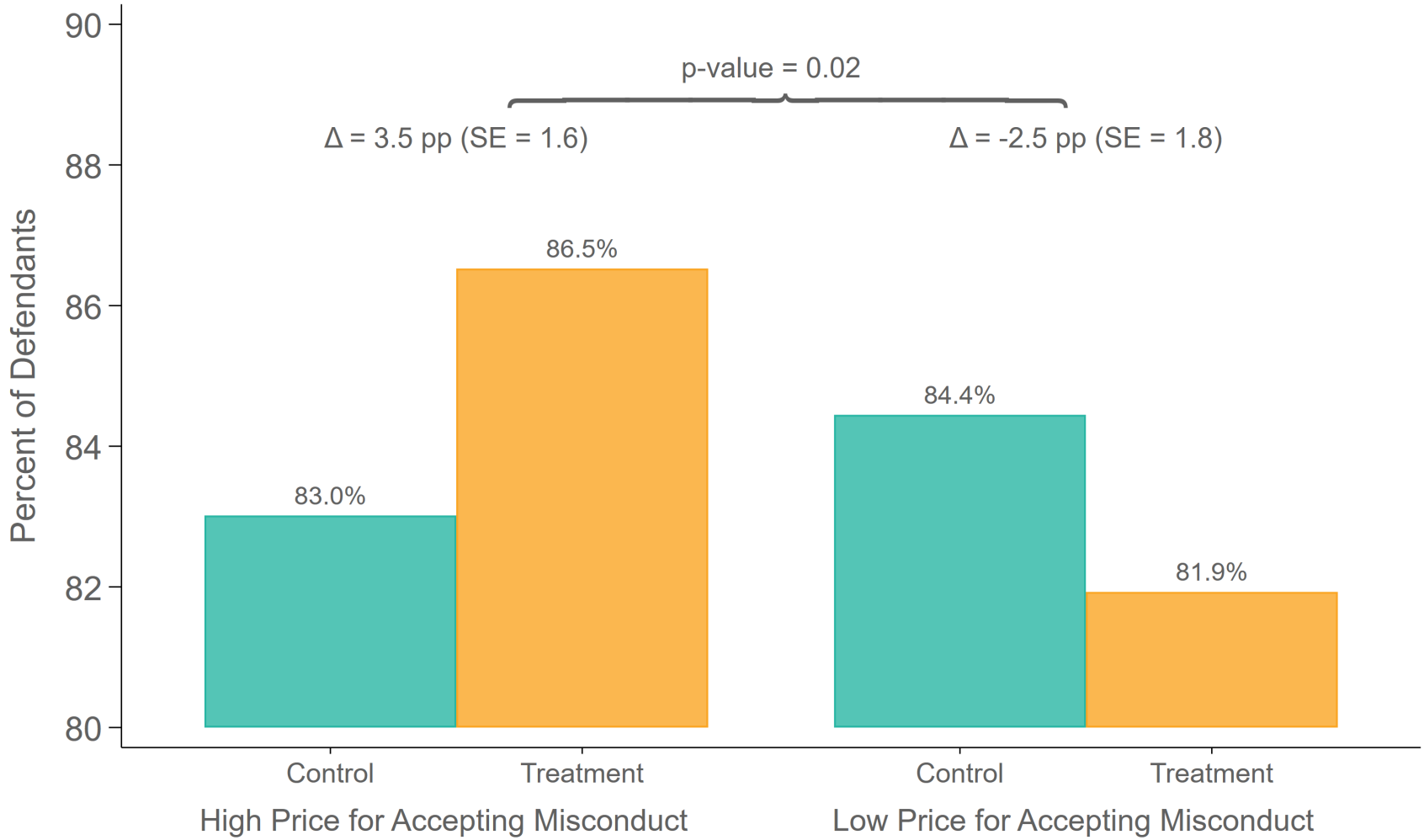
Note. This figure plots the distribution of judges by the release rate increase they require to accept a 5 percentage point increase in misconduct, relative to a baseline of 80 percent release and 20 percent misconduct. The unwilling category includes judges who never accept the higher misconduct rate even at 100 percent release.

Baseline Preferences Over Type II vs. Type I Error Costs



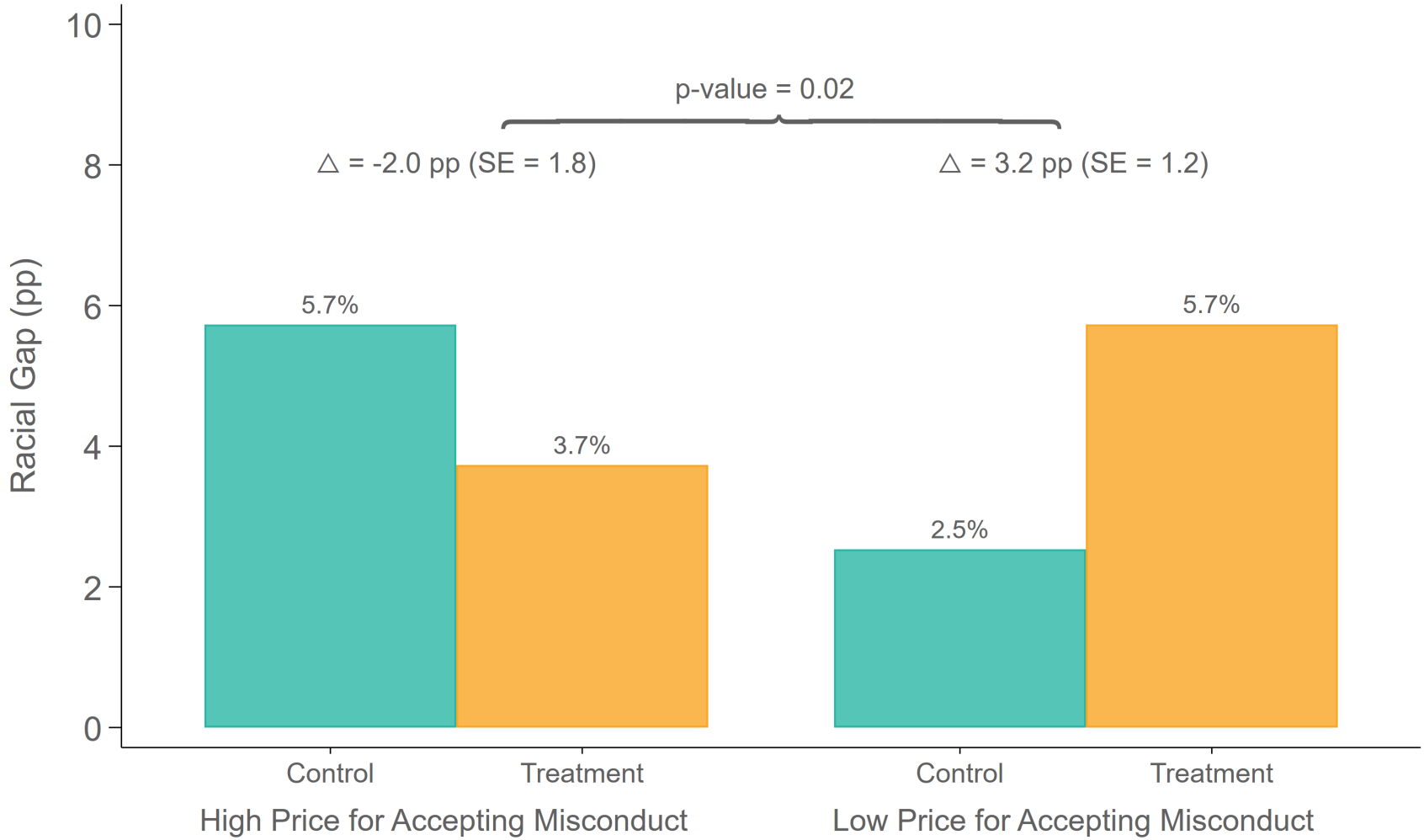
Note. This figure plots the distribution of judges by the release rate increase they require to accept a 5 percentage point increase in misconduct, relative to a baseline of 80 percent release and 20 percent misconduct. The unwilling category includes judges who never accept the higher misconduct rate even at 100 percent release.

Release Rates by Baseline Type II/Type I Error Costs



Note. This figure plots release outcomes by treatment status and judges' baseline preferences over Type II vs. Type I error costs (N=45 judges, 27k cases). Judges are classified as high price if they accept a 5 percentage point increase in misconduct only at a 100 percent release rate or never accept it, and low price otherwise. We report the control mean for low price judges; other estimates come from a regression of case-level release on treatment interacted with the high price indicator, controlling for court-by-time effects. Standard errors are clustered by judge.

Release Rate Gaps by Baseline Type II/Type I Error Costs



Note. This figure plots the racial gap in release by treatment status and judges' baseline release versus misconduct preferences (Type II vs. Type I error costs) (N=45 judges, 27k cases). Judges are classified as high price if they accept a 5 percentage point increase in misconduct only at a 100 percent release rate or never accept it, and low price otherwise. We report the control mean racial gap for low price judges; other estimates come from a regression of case-level release on treatment interacted with race and the high price indicator, controlling for court-by-time effects. Standard errors are clustered by judge.

Potential Explanations

We explore three potential explanations for our results:

✗ 1. Judges need specific guidance on each case to change their release decisions, not just court statistics and individualized feedback

✓ 2. Increasing the salience of racial disparities and/or misconduct may lead to backlash or other unintended consequences

✓ 3. Judges with high underlying preferences for release may have minimal scope for adjustment even if beliefs are corrected

Roadmap

1. Taking Theory to the Field
2. Baseline Beliefs
3. Treatment Effects on Endline Beliefs
4. Treatment Effects on Pretrial Outcomes
5. Mechanisms
6. Conclusion

Conclusion

- We study the results from two large-scale randomized trials testing interventions to correct judges' beliefs and reduce release disparities
- Treated judges update their beliefs after receiving information and training, but do not change their release rates or disparities
- Results run counter to the increasingly widespread view that beliefs drive discrimination, suggesting other interventions are needed
- Highlight the importance of considering preference- and behavioral frictions when moving from the lab to the field

Consent Form

Study Title: Reducing Racial Disparities in Bail Decisions

Researchers: Will Dobbie, Harvard University

Crystal S. Yang, Harvard University

Version Date: 2/19/2020

Key Information

The following is a short summary of this study to help you decide whether or not to be a part of this study. More detailed information is listed later on in this form.

Why am I being invited to take part in a research study?
We invite you to take part in a research study because you are a [redacted].

What should I know about a research study?
- Someone will explain this research study to you.
- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
- You can ask all the questions you want before you decide.

Why is this research being done?
We are interested in helping improve judicial decision-making and reduce racial disparities in bail decisions. Following the completion of the study, we will share our findings with both you and [redacted], as well as help implement any "best practices" that we discover during the study.

How long will the research last and what will I need to do?
We expect that you will be in this research study for 30 months. You will be asked to complete a short survey at the beginning of the study, as well as a second short survey at the end of the study. We expect each survey to take about 30-45 minutes. We may also provide you with different decision-making aids during the study, which you are free to use (or not) at your own discretion. Members of the research team may also attend your courtroom sessions and observe the decision-making process.

Is there any way being in this study could be bad for me?
The main potential risk from participating in this study is informational privacy, such as breaches of confidentiality. We perceive these risks as minor, as we intend to fully anonymize the results,

NIH Certificate of Confidentiality

DEPARTMENT OF HEALTH & HUMAN SERVICES — Public Health Service

National Institutes of Health
Bethesda, Maryland 20892
http://www.nih.gov

CERTIFICATE OF CONFIDENTIALITY

Number:
CC-OD-20-068

Issued To

Harvard University

conducting research known as

"Reducing Racial Bias in Bail Decisions"

In accordance with the provisions of section 301(d) of the Public Health Service Act, 42 U.S.C. 241(d), this Certificate is issued to the Principal Investigator, *Dr. Will Dobbie,* and the *Harvard University* to protect the privacy of subjects in the above named *single-site/single-protocol* research study, which is collecting or using identifiable, sensitive information. If there is a discrepancy between the terms used in this Certificate and section 301(d), the statutory language will control.

Research data containing identifiable, sensitive information collected during this study initiated on 01/01/2020 (and concluding on 12/31/2025) is covered by the Certificate. Identifiable, sensitive information protected by the Certificate and all copies thereof are protected for perpetuity.

The recipient of this Certificate shall comply with all requirements of subsection 301(d) of the Public Health Service Act.

This Certificate does not represent an endorsement of the research project by the Department of Health and Human Services. Information collected during the term of the Certificate is protected in perpetuity. However, this Certificate does not protect information collected from participants enrolled after the term of the Certificate.

05/16/2024
Date

Lynda L Tall

NIH Certificates of Confidentiality Staff
Office of Extramural Research
National Institutes of Health

The expiration date on this certificate reflects an automatic one-year extension granted by NIH for CoCs active as of April 15, 2020 as a result of the COVID-19 public health emergency.

Representativeness at First Research Site

	Sample Mean	Non-Sample Mean	*p*-value
Release for All Defendants	87.5%	85.5%	0.687
Release for White Defendants	90.5%	89.2%	0.612
Release for Non-White Defendants	85.4%	82.3%	0.710
FTA for All Defendants	8.8%	8.9%	0.288
FTA for White Defendants	9.7%	8.4%	0.111
FTA for Non-White Defendants	8.1%	9.3%	0.799
NCA for All Defendants	11.7%	11.0%	0.201
NCA for White Defendants	11.3%	11.1%	0.519
NCA for Non-White Defendants	12.0%	10.9%	0.382
Cases	12,795	2,421	
Judges	37	14	

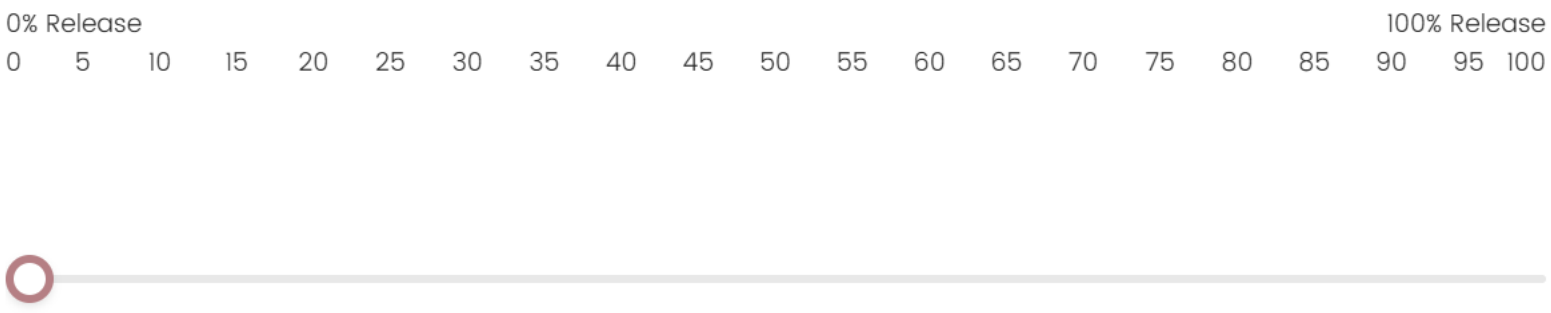
Note. This table reports representativeness at baseline at our first research site.

Instructions for Release Questions

We're interested in learning about how often you grant **pretrial release (defined as release on recognizance, release with non-financial conditions, or release on financial conditions/cash/money bail)**. For example, if you grant pretrial release for 70 out of every 100 defendants, you should move the slider to 70%. Please provide your best guess.

Release Questions

Over the past year, what was your **pretrial release rate** for **all** defendants?



Release Questions

Over the past year, what was your **pretrial release rate** for **white** defendants?

0% Release — 100% Release

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Over the past year, what was your **pretrial release rate** for **non-white** defendants (e.g. black and Hispanic)?

0% Release — 100% Release

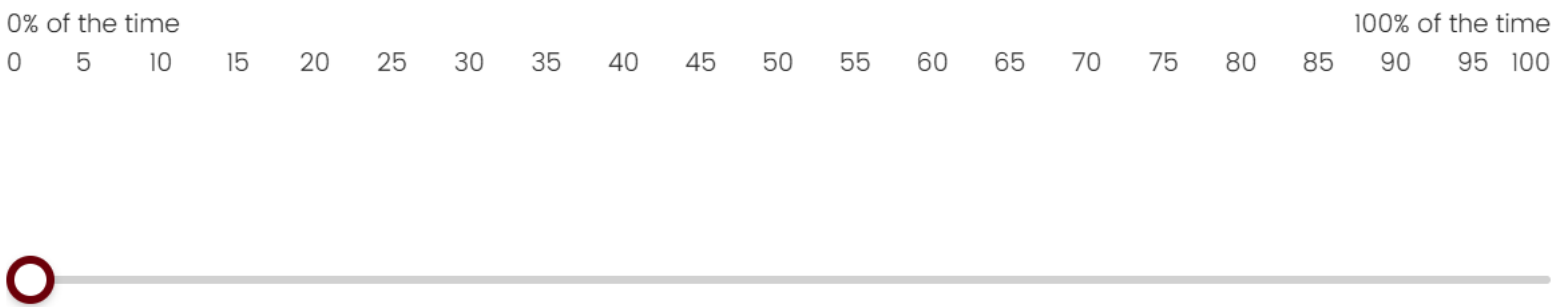
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Instructions for FTA Questions

In the next set of questions, we're interested in learning about how often the defendants you released **failed to appear (FTA)** for a required court appearance. For example, if 10 out of every 100 defendants you released FTA'ed, you should move the slider to 10%. Please provide your best guess.

FTA Questions

Over the past year, about how often did **all** defendants that you released **FTA**?



FTA Questions

Over the past year, about how often did the **white** defendants that you released **FTA**?



Over the past year, about how often did the **non-white** defendants (e.g. black and Hispanic) that you released **FTA**?



Eliciting Baseline Preferences

We want to know how much higher your release rate would need to be for you to be willing to accept a higher pretrial misconduct rate. Specifically, let's suppose that your starting point is a release rate of 80% and a pretrial Y ∈ {FTA, NCA} rate of 20%.

At what release rate $X \in \{85, 90, 95, 100\}$ would you be willing to accept a pretrial Y ∈ {FTA, NCA} rate of 25%?

Please indicate which of the two options you would prefer.

Release Rate of 80% AND Y Rate of 20%	Release Rate of X% AND Y Rate of 25%
○	○

IAT Questions

In the task that follows, a picture or word will appear in the middle of the screen. Click the category on the left for items that belong to the category “Black” or "Unpleasant" and click the category on the right for items that belong to the category “White” or "Pleasant".

Items will appear one at a time. If you make a mistake, a red X will appear. Click the other category to continue.

Go as fast as you can while being accurate.

CLICK HERE TO CONTINUE

BLACK
or
Unpleasant

WHITE
or
Pleasant

IAT Questions

If you make a mistake, a **X** will appear.

Please correct by clicking on the other category.



BLACK
or
Unpleasant

WHITE
or
Pleasant

IAT Questions

If you make a mistake, a **X** will appear.

Please correct by clicking on the other category.

Hatred

BLACK
or
Pleasant

WHITE
or
Unpleasant

Court Release Questions

In this question we are interested in the **differences in pretrial release rates** (defined as release on recognizance, release with non-financial conditions, or release on financial conditions/cash/money bail) in your court system. In your court system, how do you think pre-trial release rates compare for white and non-white (e.g. black and Hispanic) defendants?

- (●) Pretrial release rates are **higher** for **white defendants** than for non-white defendants
- ( ) Pretrial release rates are **higher** for **non-white defendants** than for white defendants
- ( ) Pretrial release rates are **equal** for white and non-white defendants

Judge Demographics

	Control Mean	Treatment Mean	p-value
Age	51.7	54.5	0.524
Male	69.6%	72.0%	0.896
White	76.0%	76.0%	0.913
Law Degree (JD)	58.8%	66.7%	0.273
Years of Experience	11.5	10.7	0.682
Judges	23	25	48

Note. This table reports demographics for the 48 judges in our analysis sample who took the baseline survey.

Baseline Case Outcomes

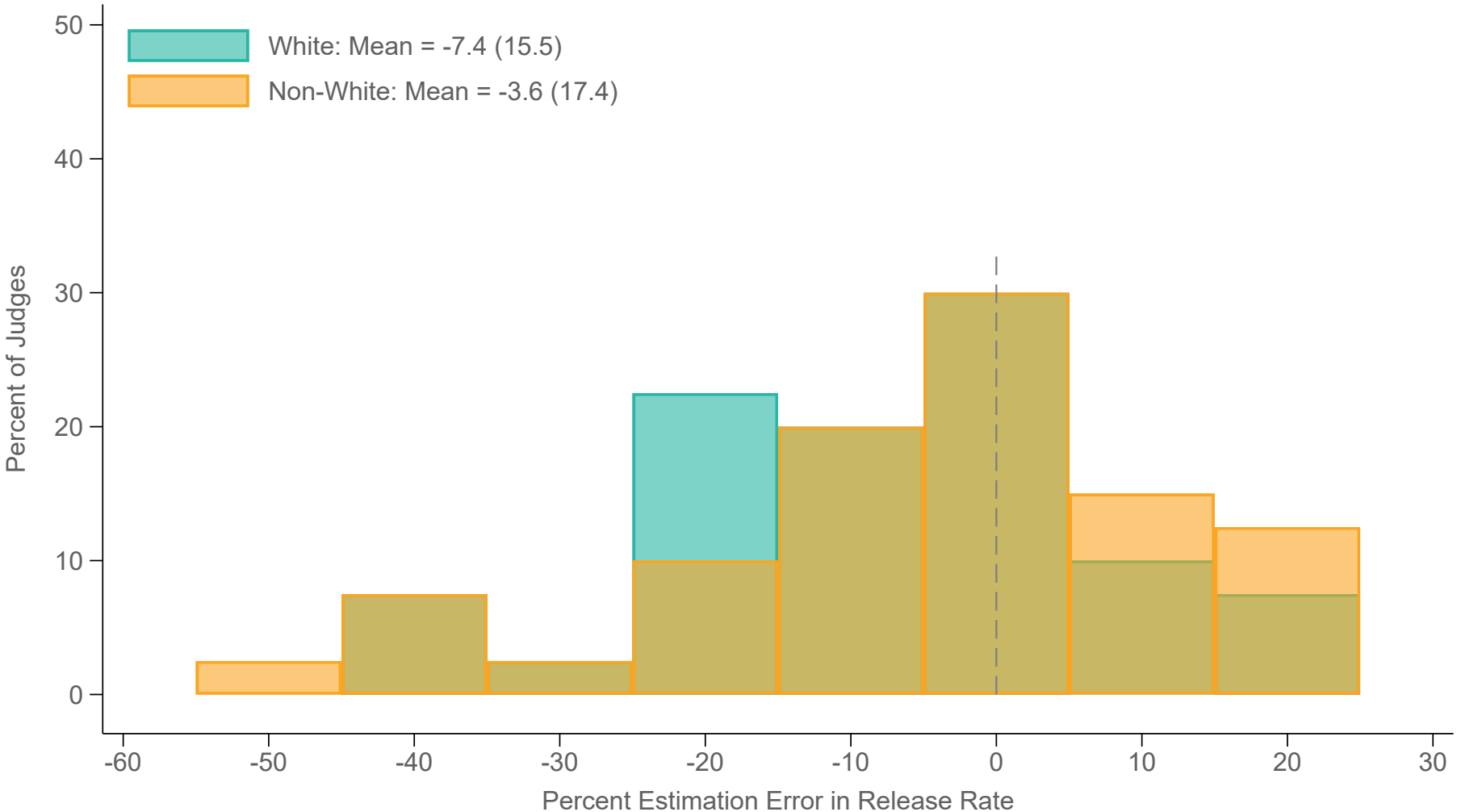
	Control Mean	Treatment Mean	p-value
Release for All Defendants	86.0%	85.9%	0.623
Release for White Defendants	88.3%	88.4%	0.706
Release for Non-White Defendants	84.1%	83.9%	0.586
FTA for All Defendants	10.9%	10.9%	0.552
FTA for White Defendants	11.4%	11.7%	0.544
FTA for Non-White Defendants	10.5%	10.3%	0.671
NCA for All Defendants	9.4%	10.9%	0.064
NCA for White Defendants	8.6%	10.3%	0.195
NCA for Non-White Defendants	10.0%	11.4%	0.148
Cases	8,515	9,610	18,125

Note. This table reports baseline case outcomes for the 51 judges in our analysis sample.

Feedback from the Treated Judges at Endline

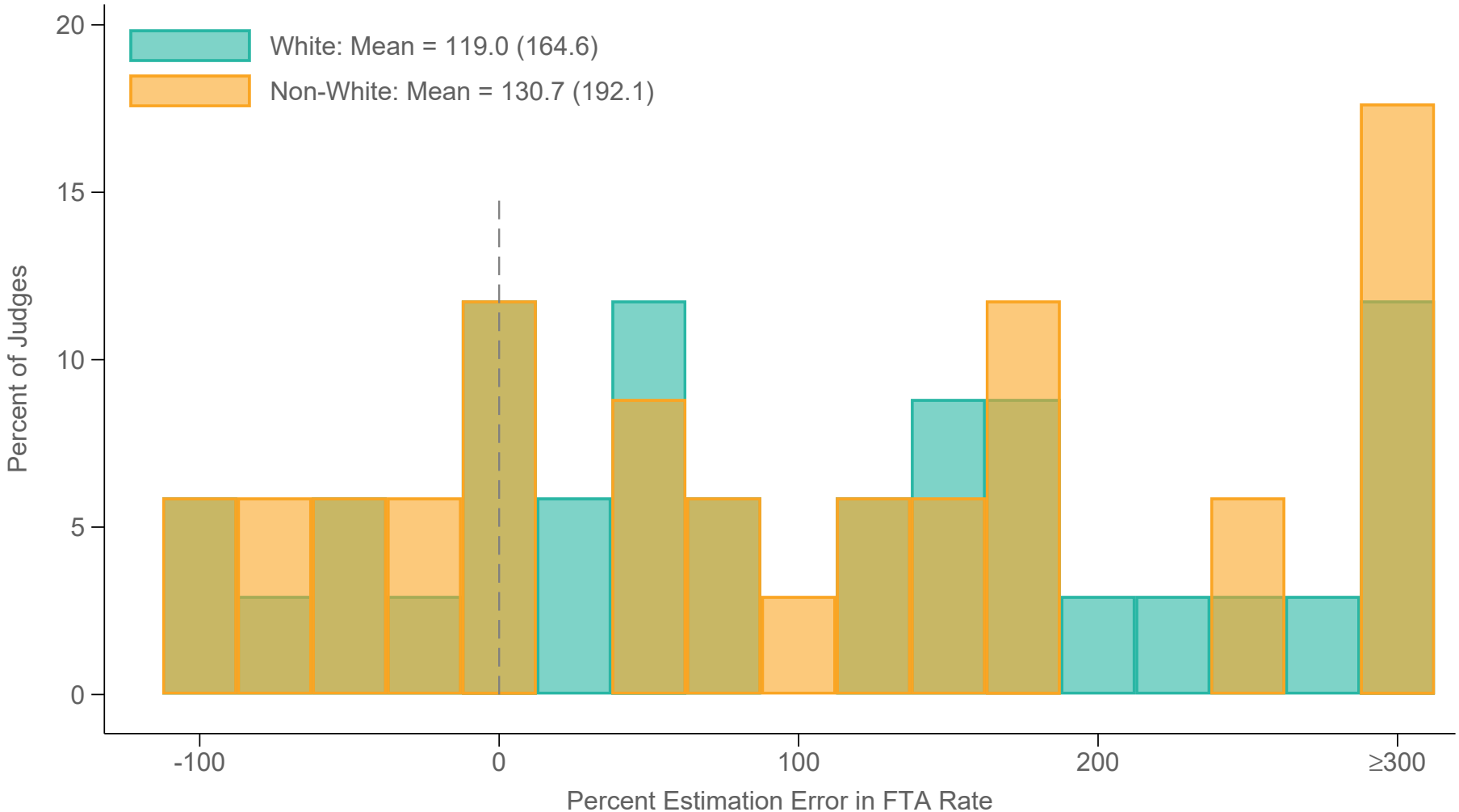
- *"Feedback on actual data is immensely helpful to understand if my perception matches reality."*
- *"Feedback was very helpful particularly the race gaps."*
- *"This was really a wonderful opportunity for me, especially as a newer judge, to learn about the impact of our decisions. It was also really helpful to see that my decisions about release seem to be on the mark for the most part (high return and arrest-free rates) because we usually only find out about the folks who we let out who violate conditions/fta."*

Estimation Error for Release Rates by Race



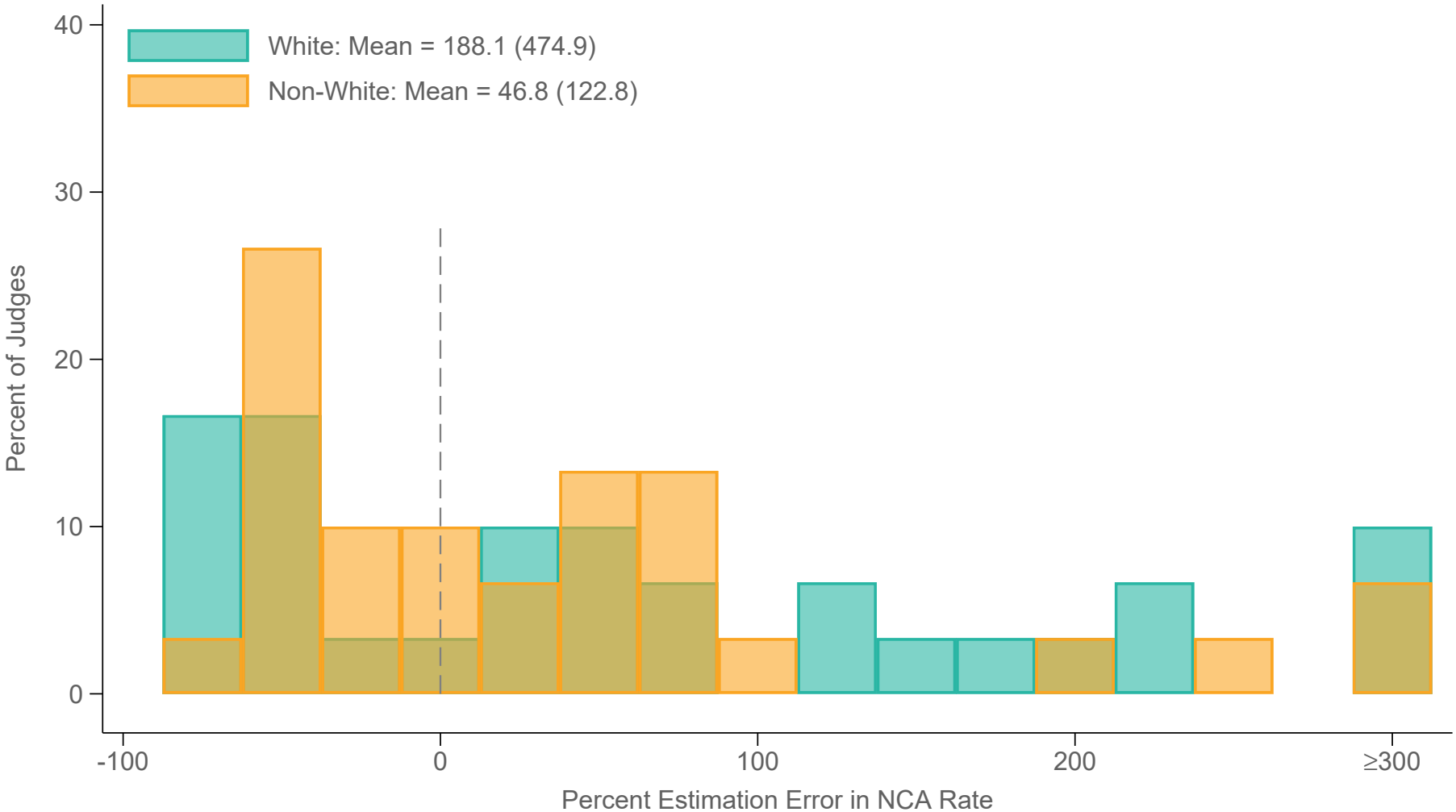
Note. This figure plots the estimation error in release rates by defendant race, defined as the percentage difference between baseline beliefs and the objective outcome (N=40). Eight judges coded as "I don't know."

Estimation Error for FTA Rates by Race



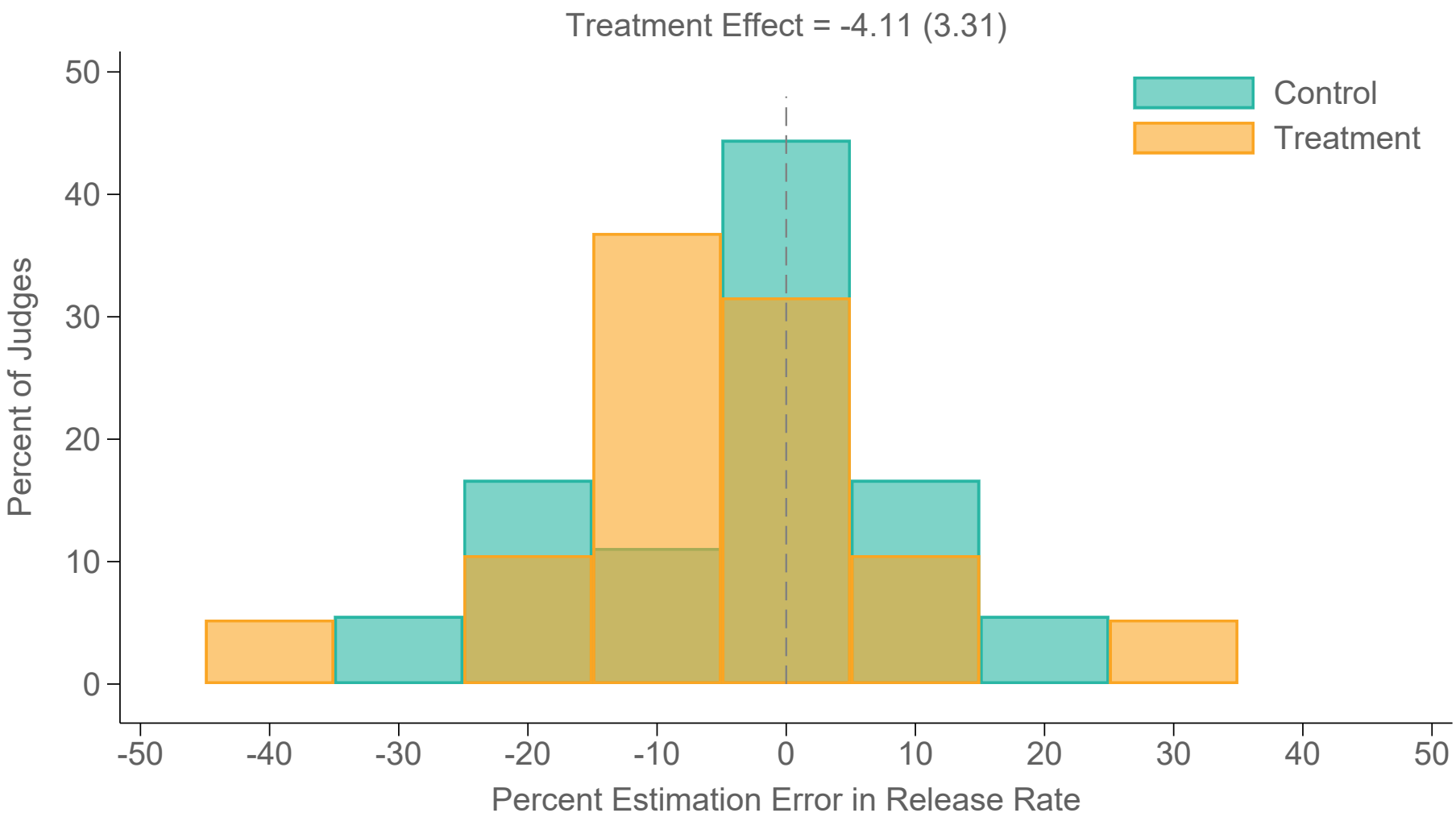
Note. This figure plots the estimation error in FTA rates by defendant race, defined as the percentage difference between baseline beliefs and the objective outcome (N=34). Fourteen judges coded as "I don't know."

Estimation Error for NCA Rates by Race



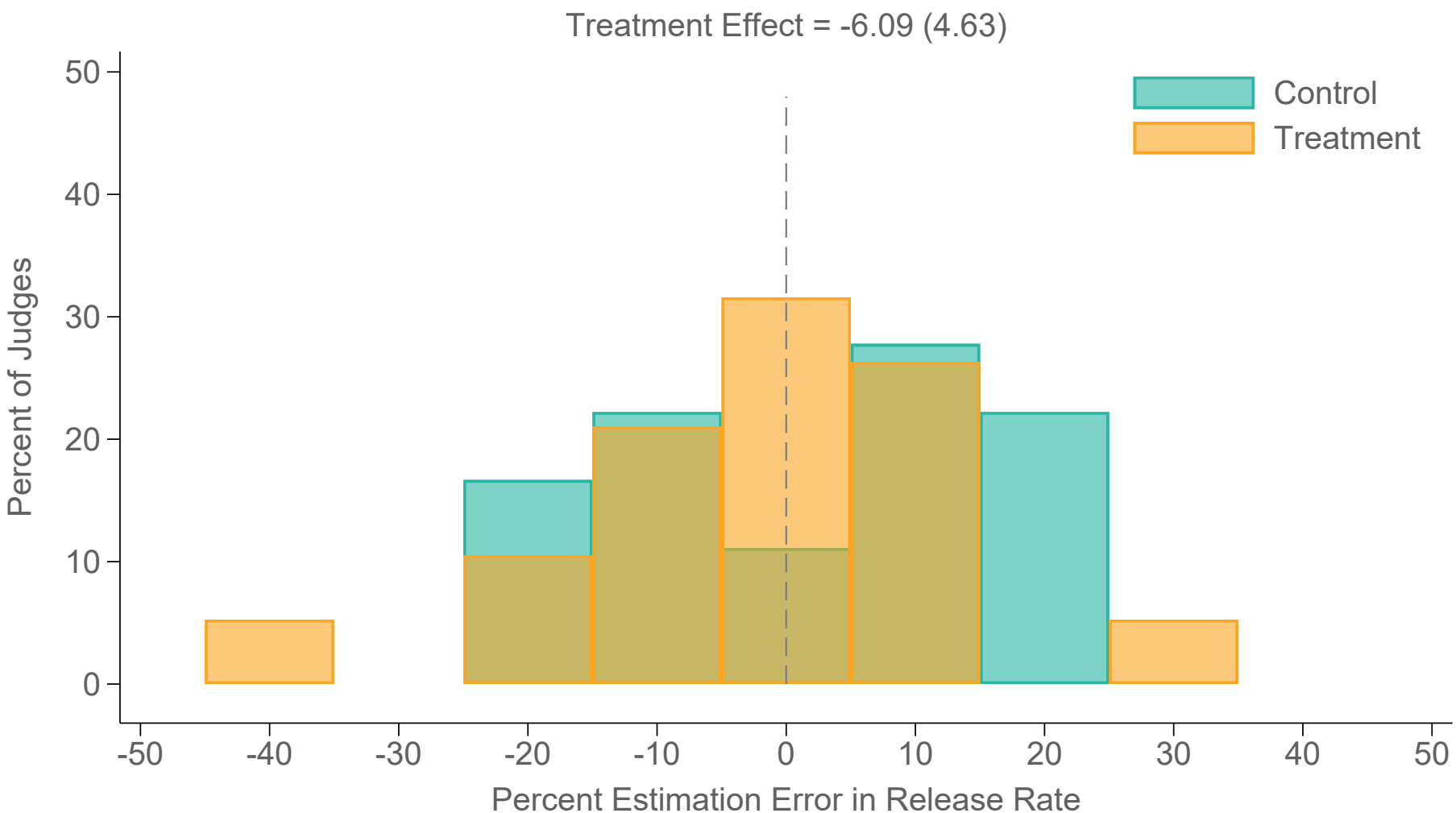
Note. This figure plots the estimation error in NCA rates by defendant race, defined as the percentage difference between baseline beliefs and the objective outcome (N=30). Eighteen judges coded as "I don't know."

Estimation Error for Endline Release Beliefs: White



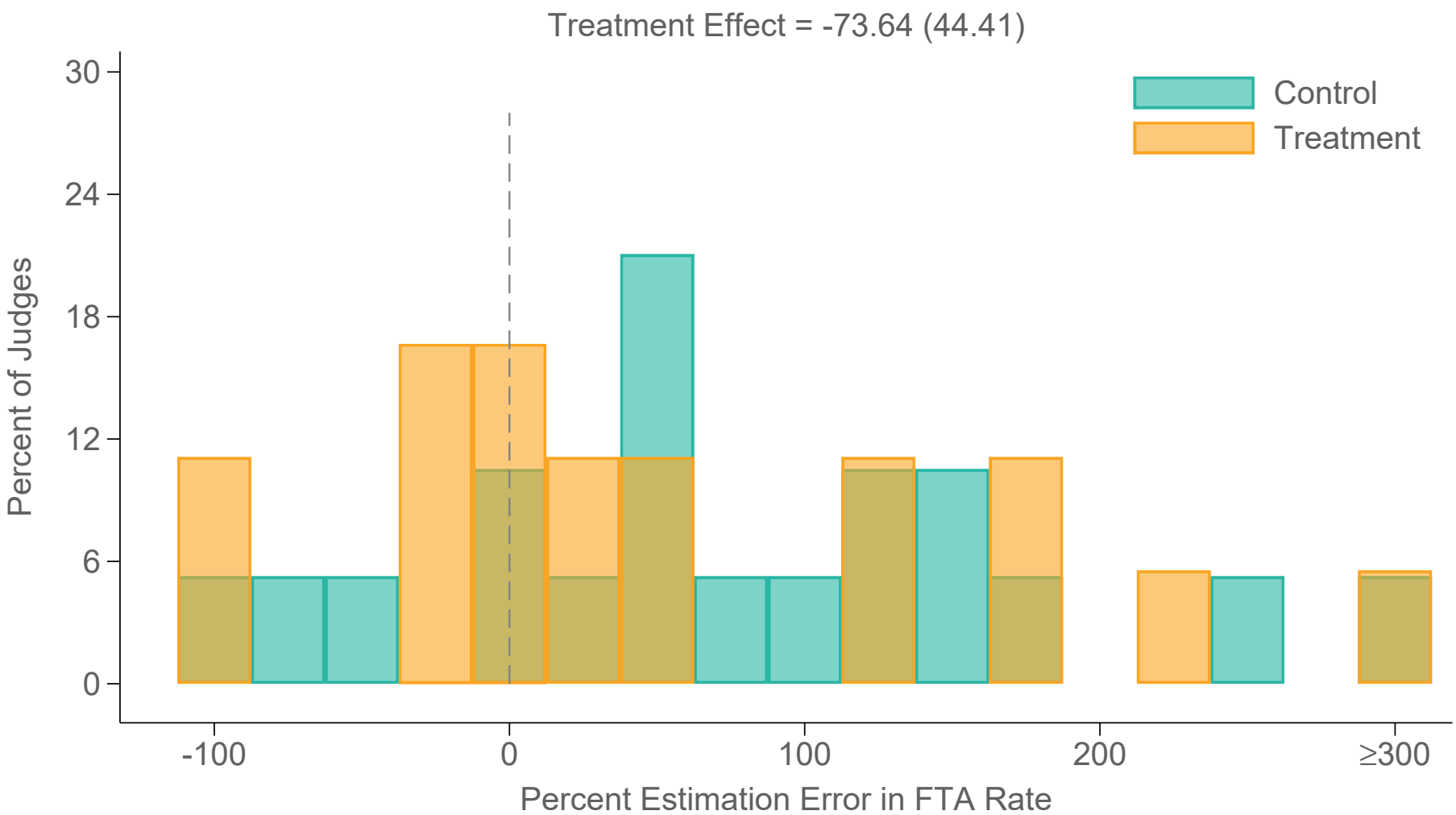
Note. This figure plots the estimation error in release rates by treatment status for White defendants, defined as the percentage difference between baseline beliefs and the objective outcome (N=37). Three judges coded as "I don't know." Treatment effects are calculated from a regression of endline beliefs on treatment status, baseline control, and strata fixed effects.

Estimation Error for Endline Release Beliefs: Non-White



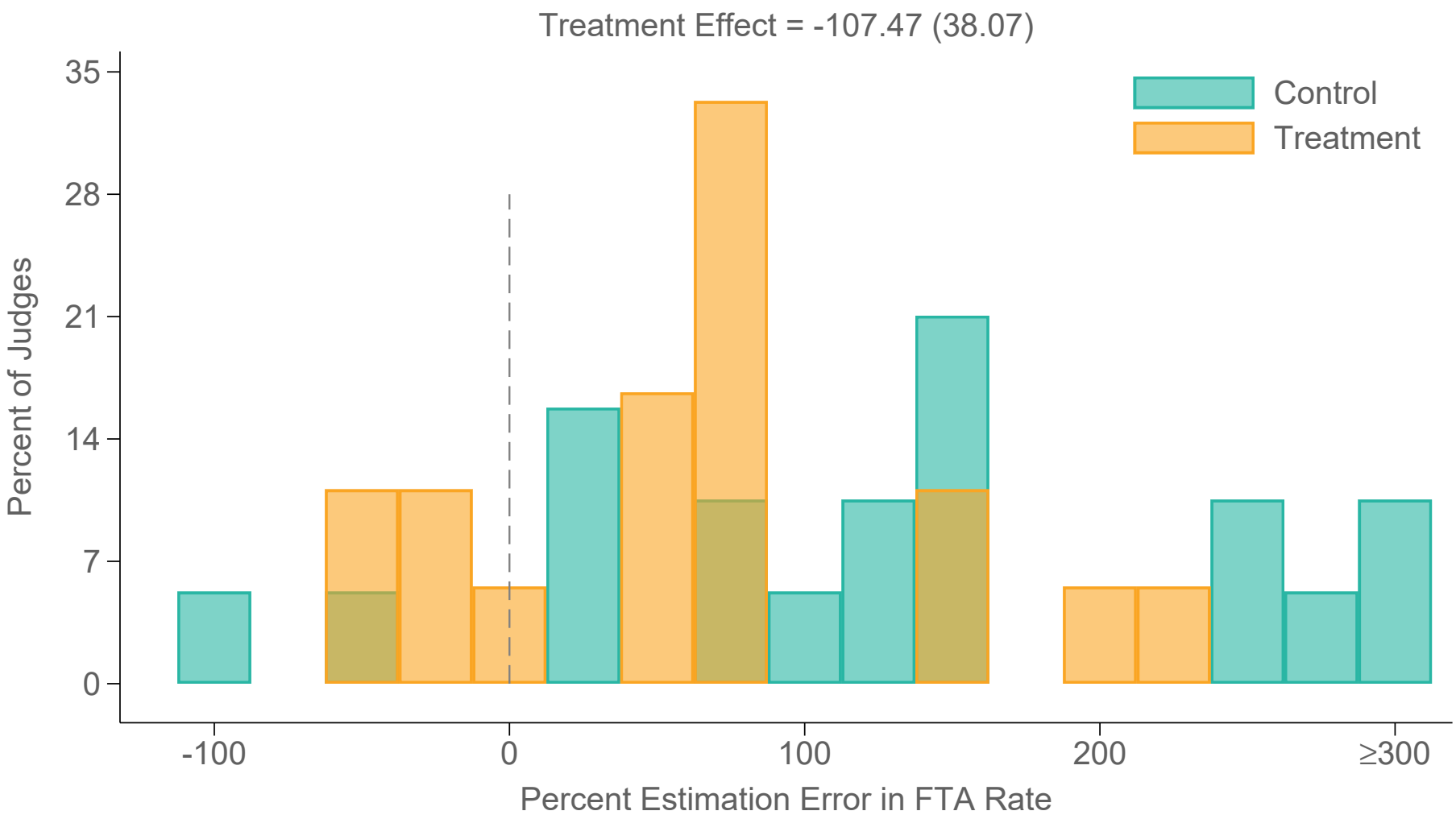
Note. This figure plots the estimation error in release rates by treatment status for Non-White defendants, defined as the percentage difference between baseline beliefs and the objective outcome (N=37). Three judges coded as "I don't know." Treatment effects are calculated from a regression of endline beliefs on treatment status, baseline control, and strata fixed effects.

Estimation Error for Endline FTA Beliefs: White



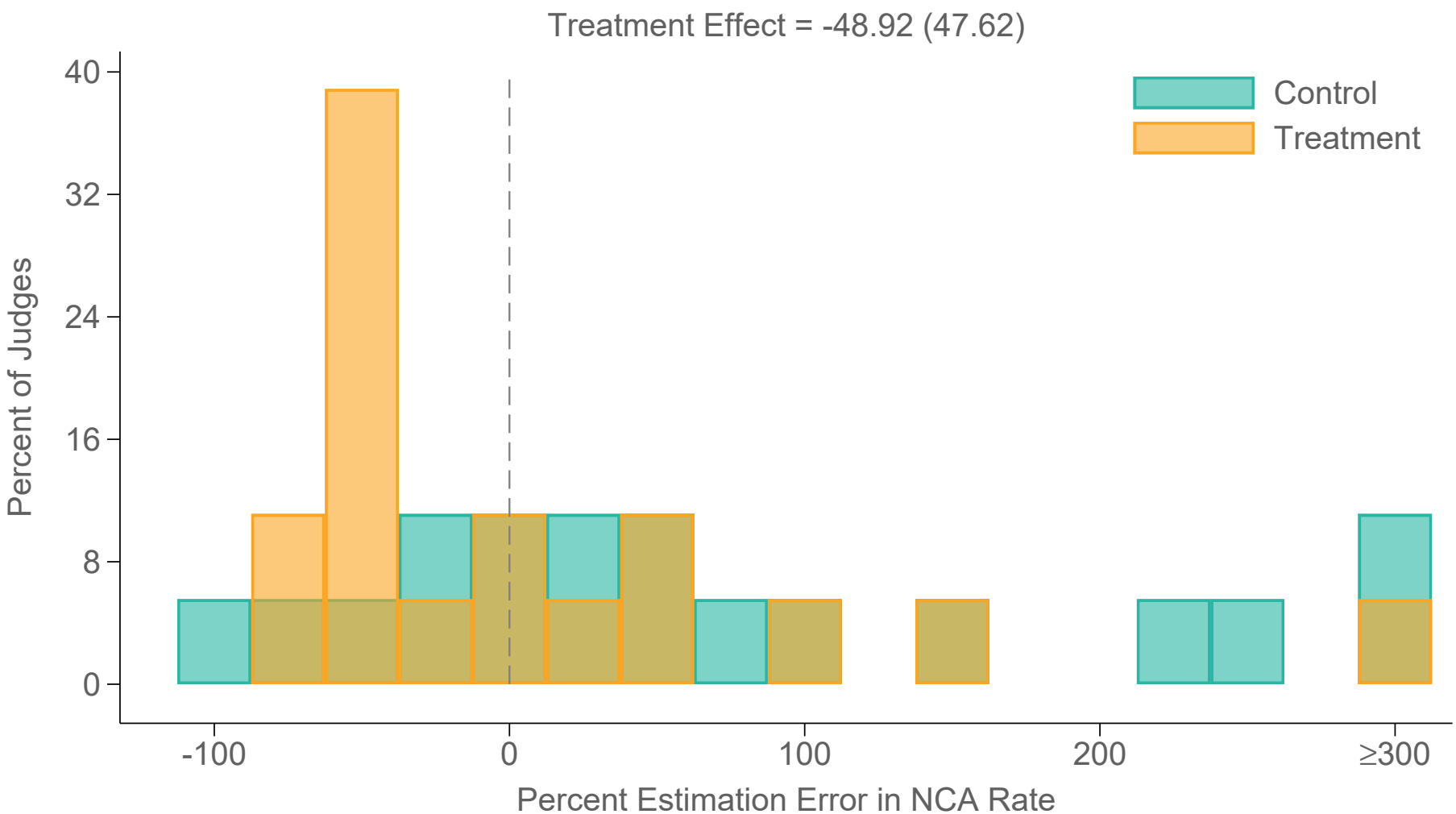
Note. This figure plots the estimation error in FTA rates by treatment status for White defendants, defined as the percentage difference between baseline beliefs and the objective outcome (N=37). Three judges coded as "I don't know." Treatment effects are calculated from a regression of endline beliefs on treatment status, baseline control, and strata fixed effects.

Estimation Error for Endline FTA Beliefs: Non-White



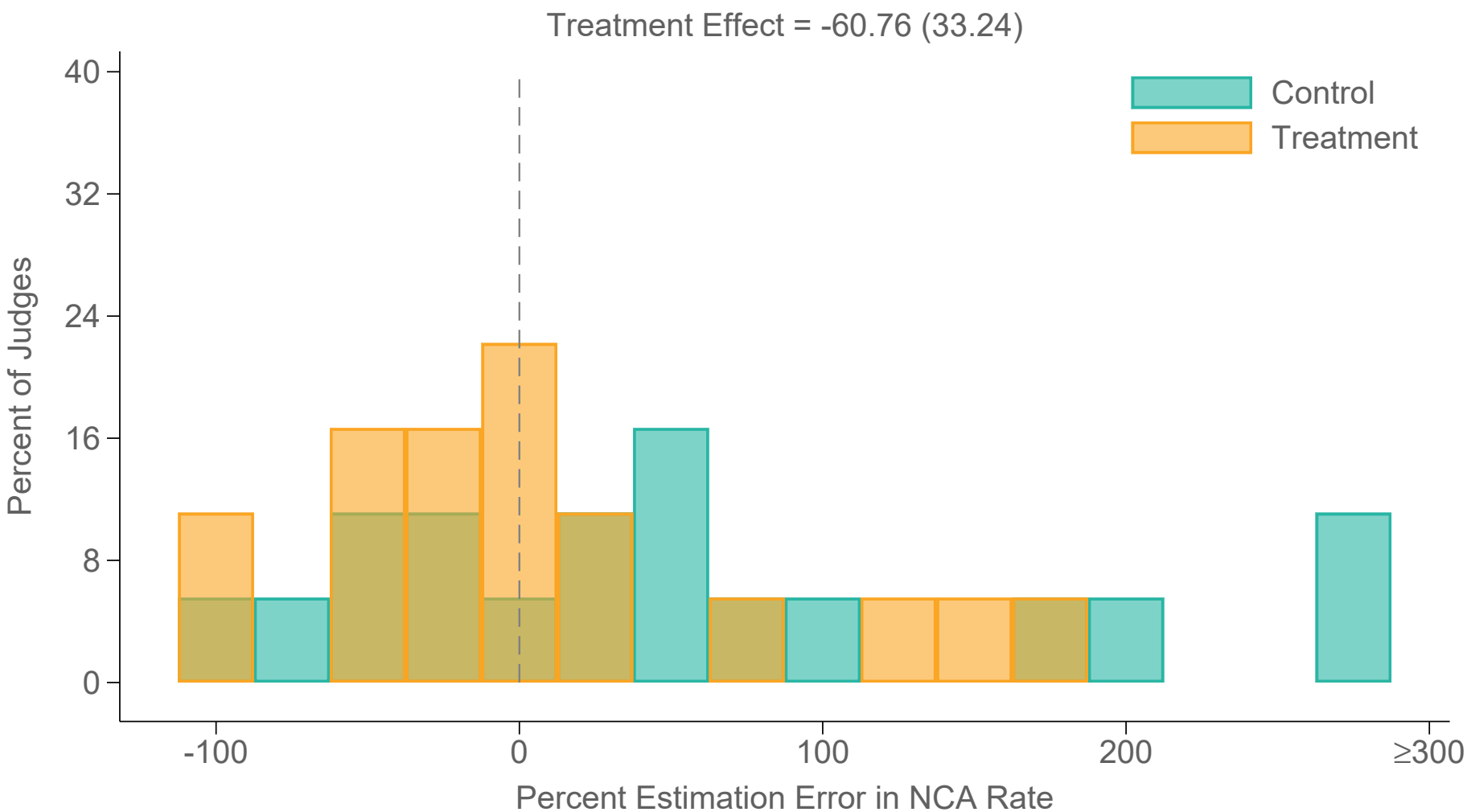
Note. This figure plots the estimation error in FTA rates by treatment status for Non-White defendants, defined as the percentage difference between baseline beliefs and the objective outcome (N=37). Three judges coded as "I don't know." Treatment effects are calculated from a regression of endline beliefs on treatment status, baseline control, and strata fixed effects.

Estimation Error for Endline NCA Beliefs: White



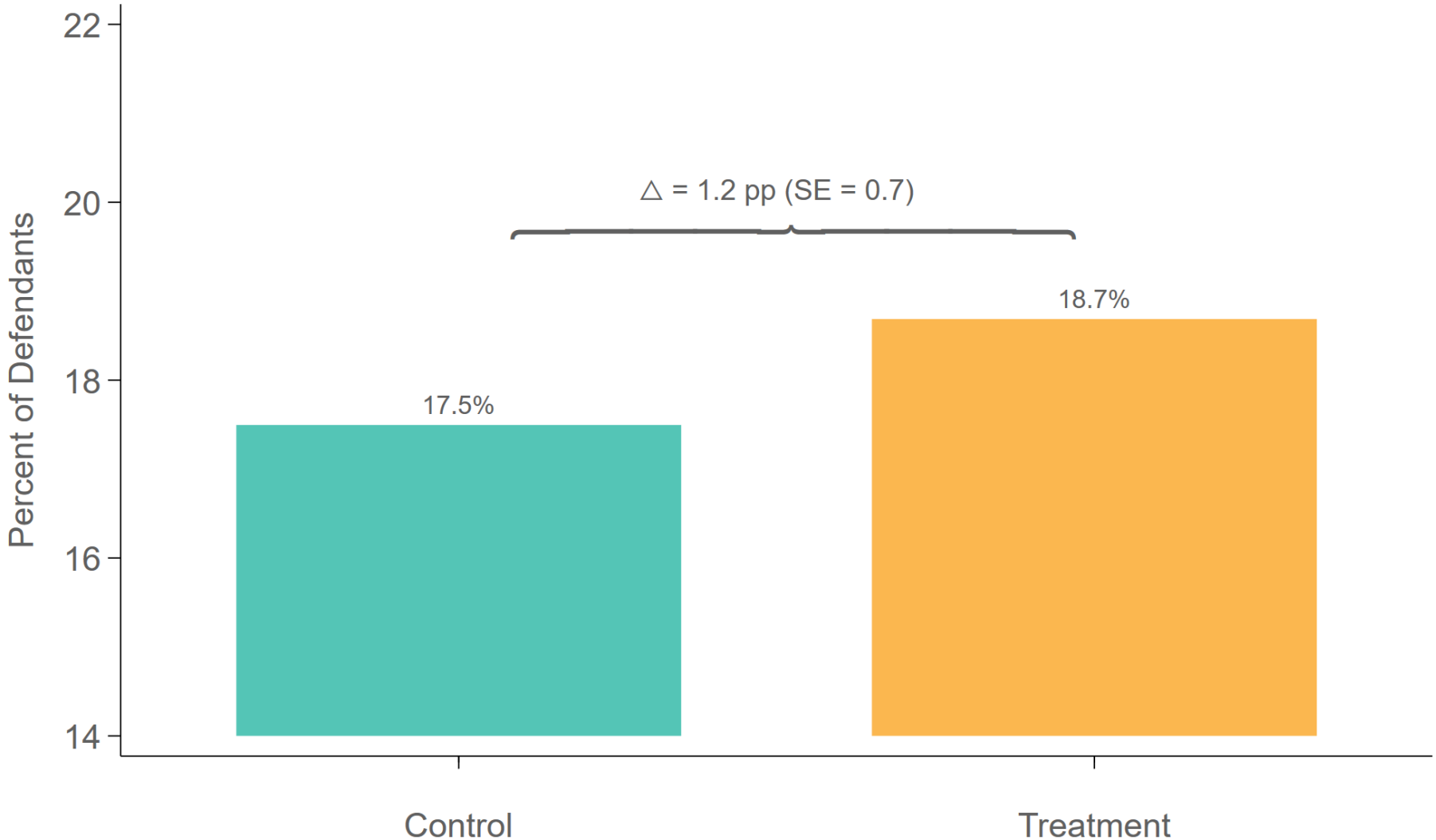
Note. This figure plots the estimation error in NCA rates by treatment status for White defendants, defined as the percentage difference between baseline beliefs and the objective outcome (N=36). Four judges coded as "I don't know." Treatment effects are calculated from a regression of endline beliefs on treatment status, baseline control, and strata fixed effects.

Estimation Error for Endline NCA Beliefs: Non-White



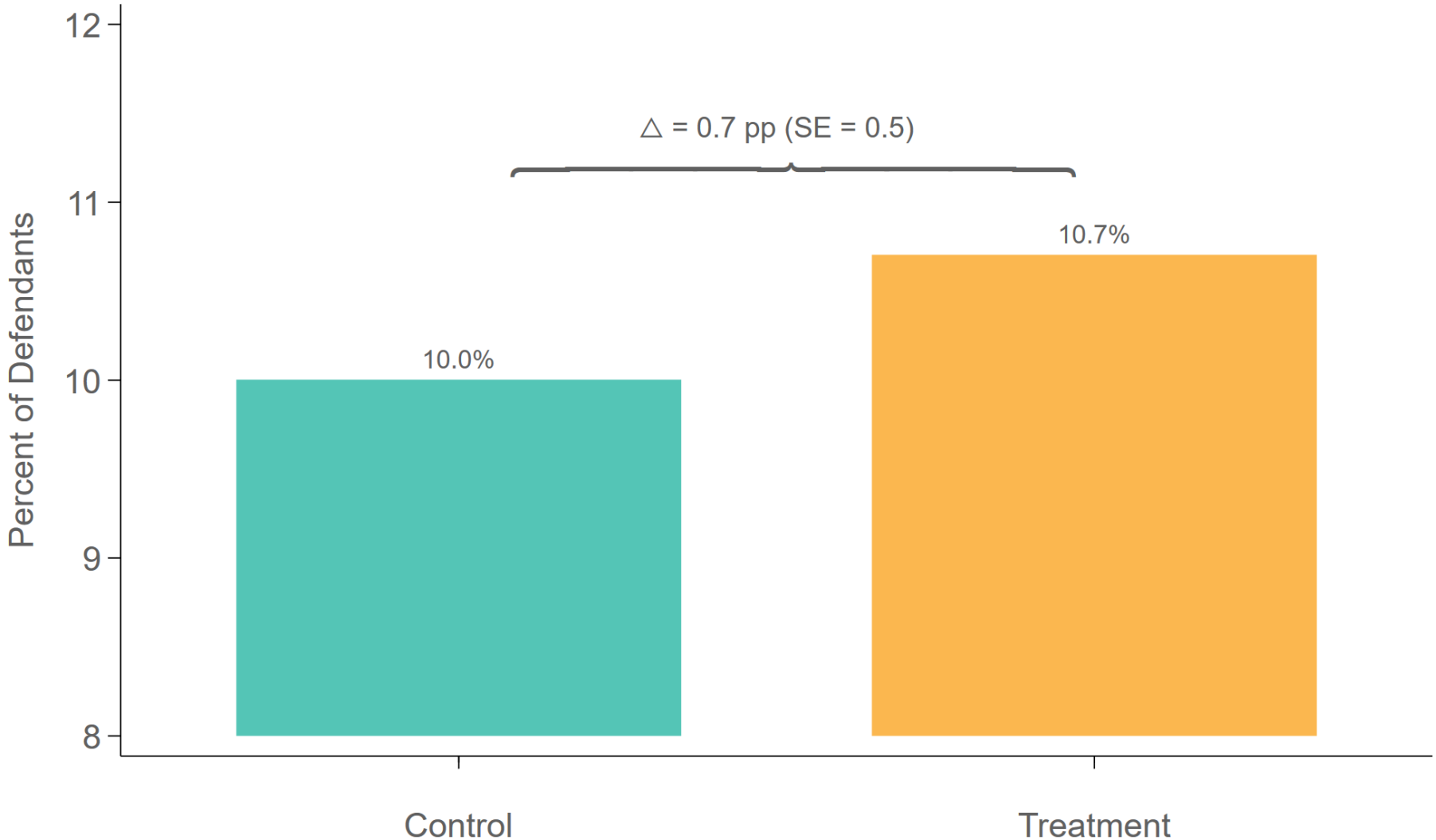
Note. This figure plots the estimation error in NCA rates by treatment status for Non-White defendants, defined as the percentage difference between baseline beliefs and the objective outcome (N=36). Four judges coded as "I don't know." Treatment effects are calculated from a regression of endline beliefs on treatment status, baseline control, and strata fixed effects.

Misconduct Rates



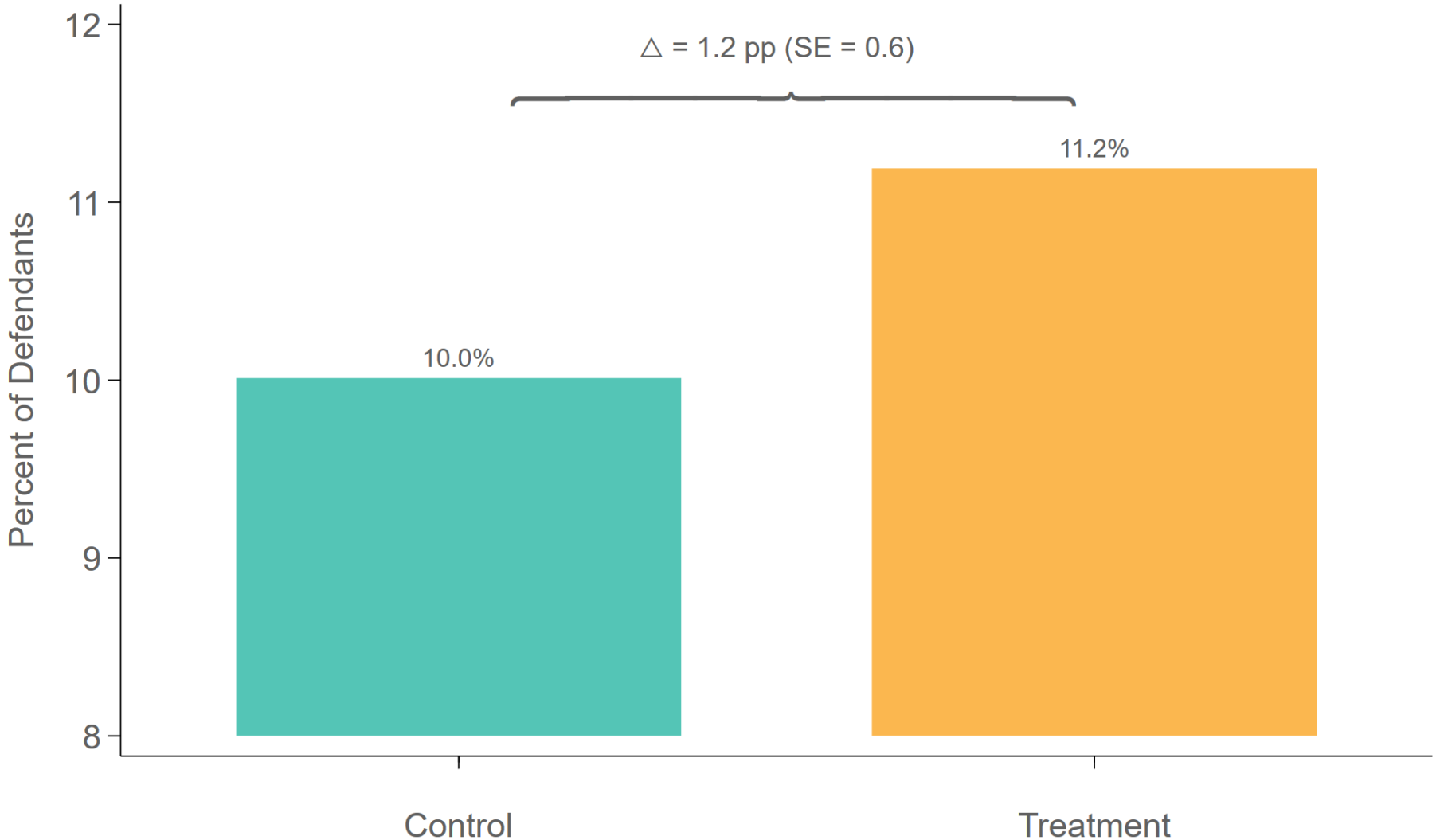
Note. This figure plots any misconduct (FTA or NCA) outcomes by treatment status (N=51 judges, 25k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

FTA Rates



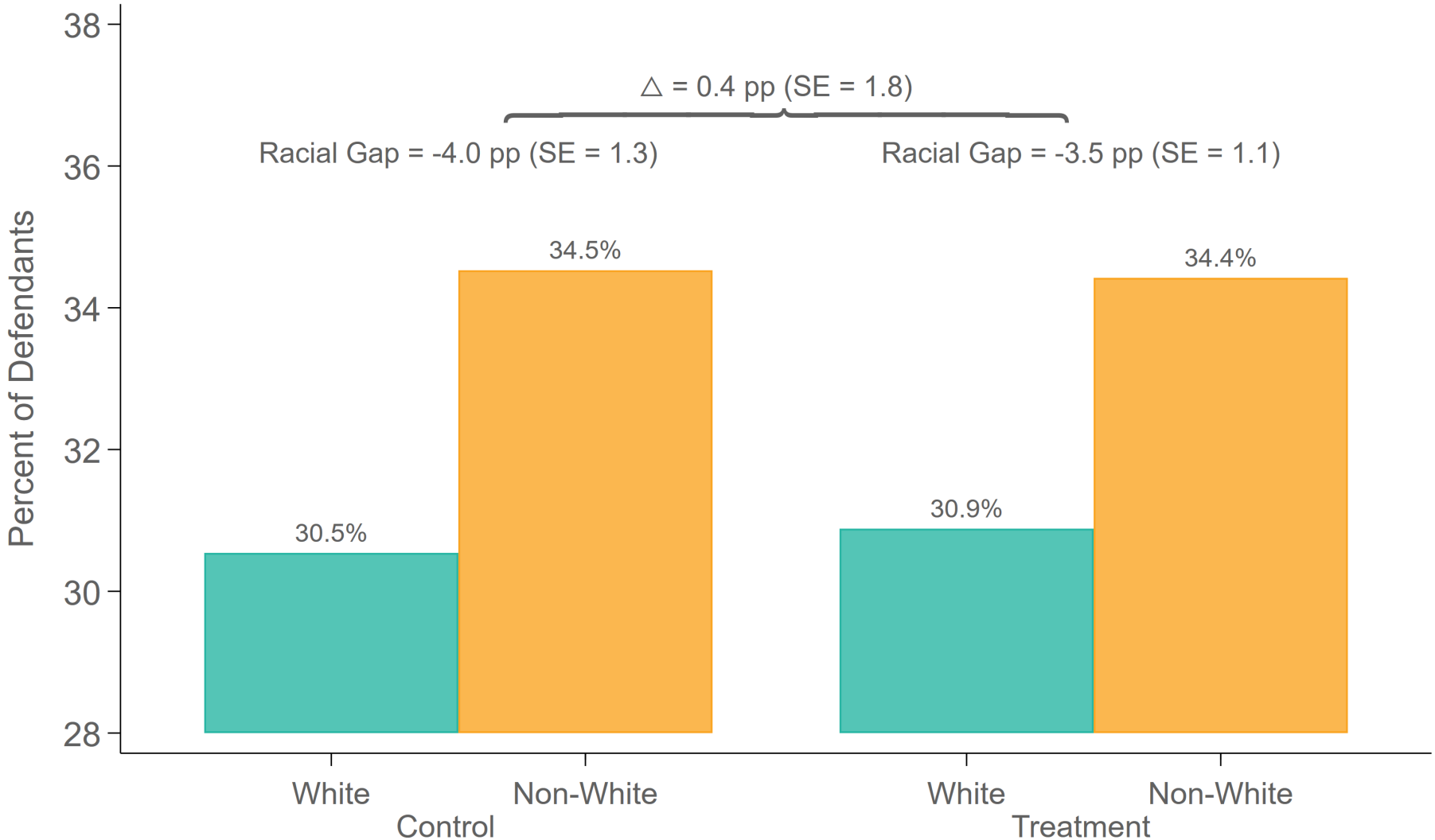
Note. This figure plots failure to appear (FTA) outcomes by treatment status (N=51 judges, 25k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

NCA Rates



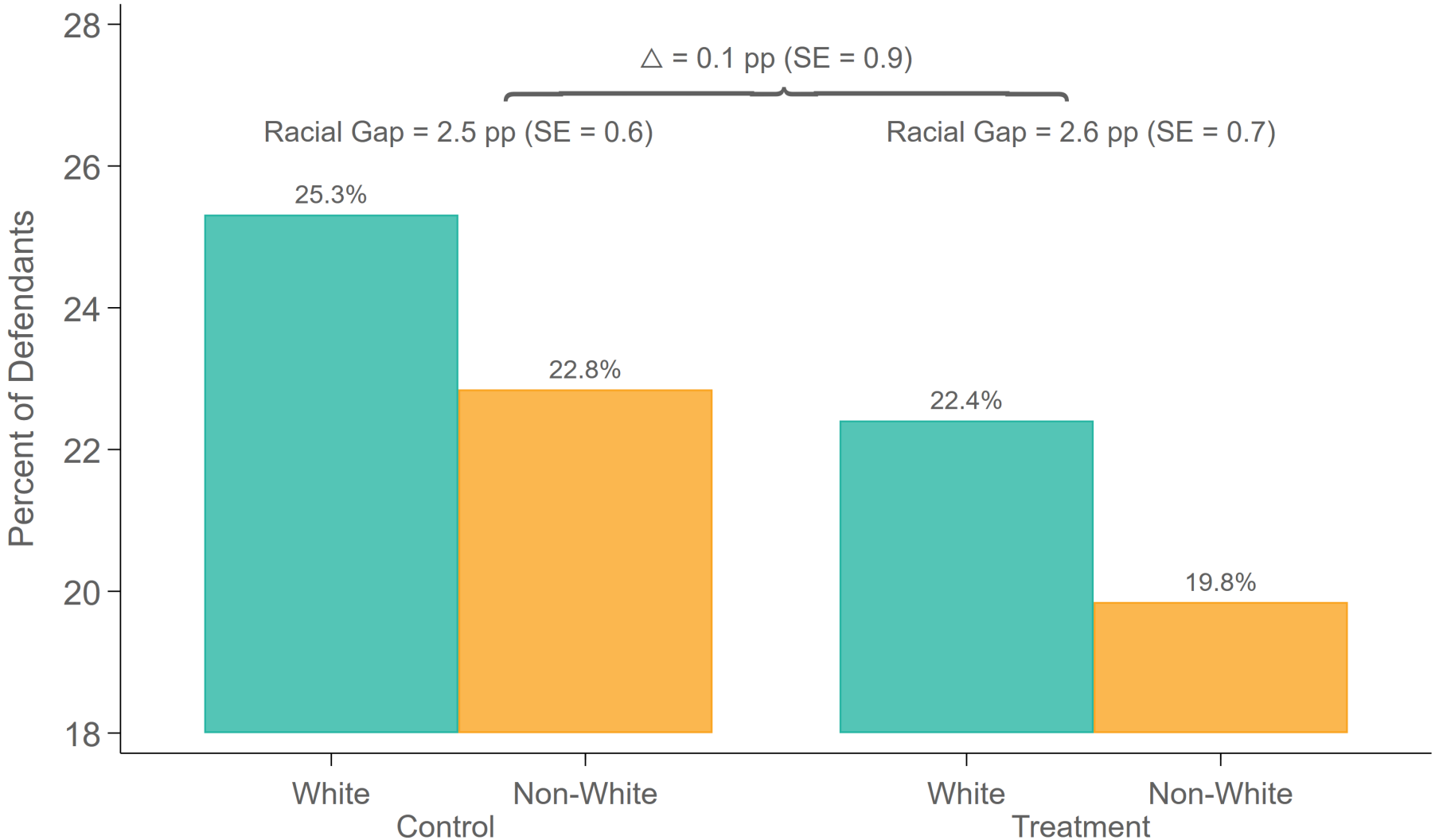
Note. This figure plots new criminal activity (NCA) outcomes by treatment status (N=51 judges, 25k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Monetary Bail Rates



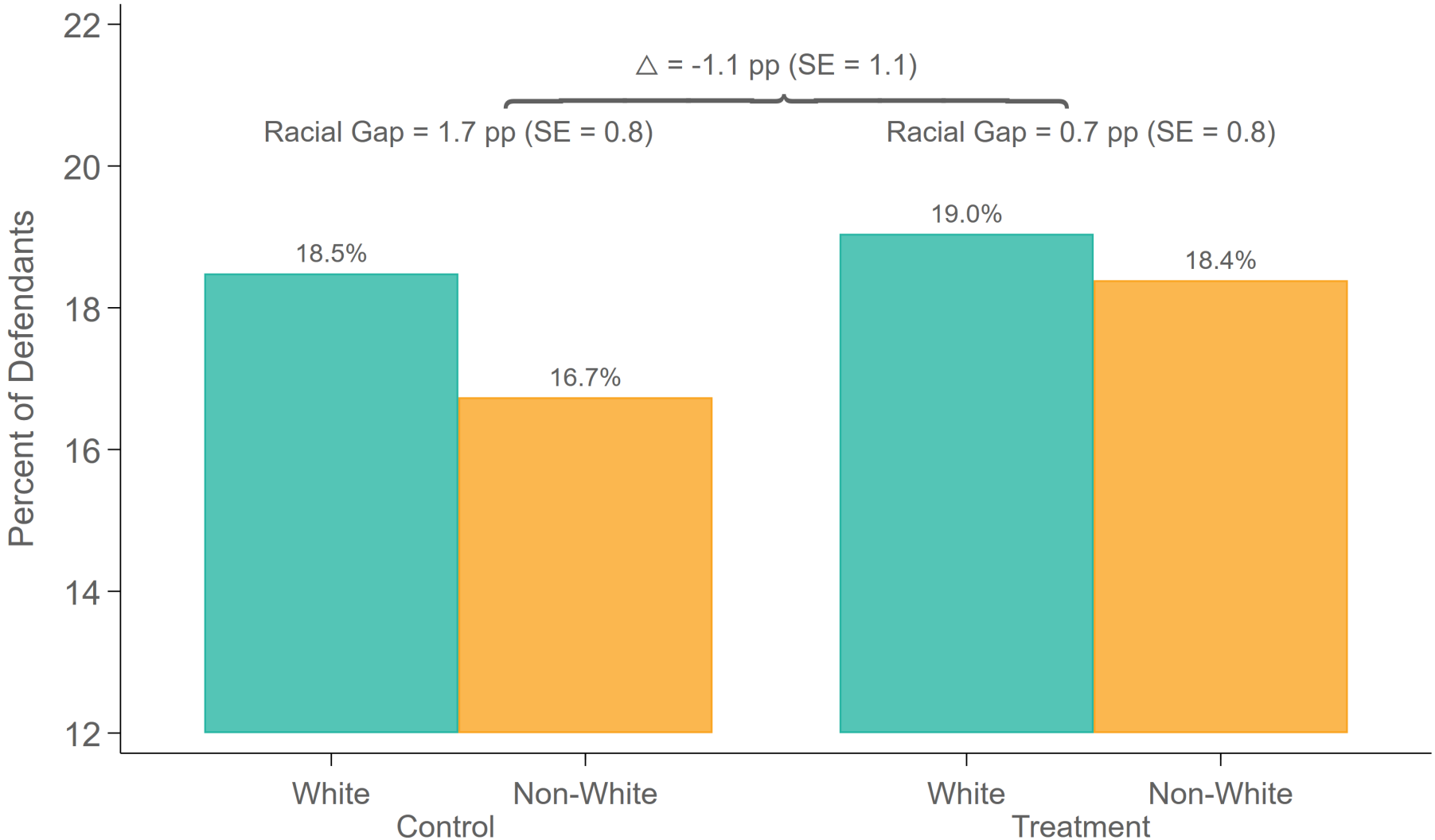
Note. This figure plots monetary bail by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Release on Recognizance Rates



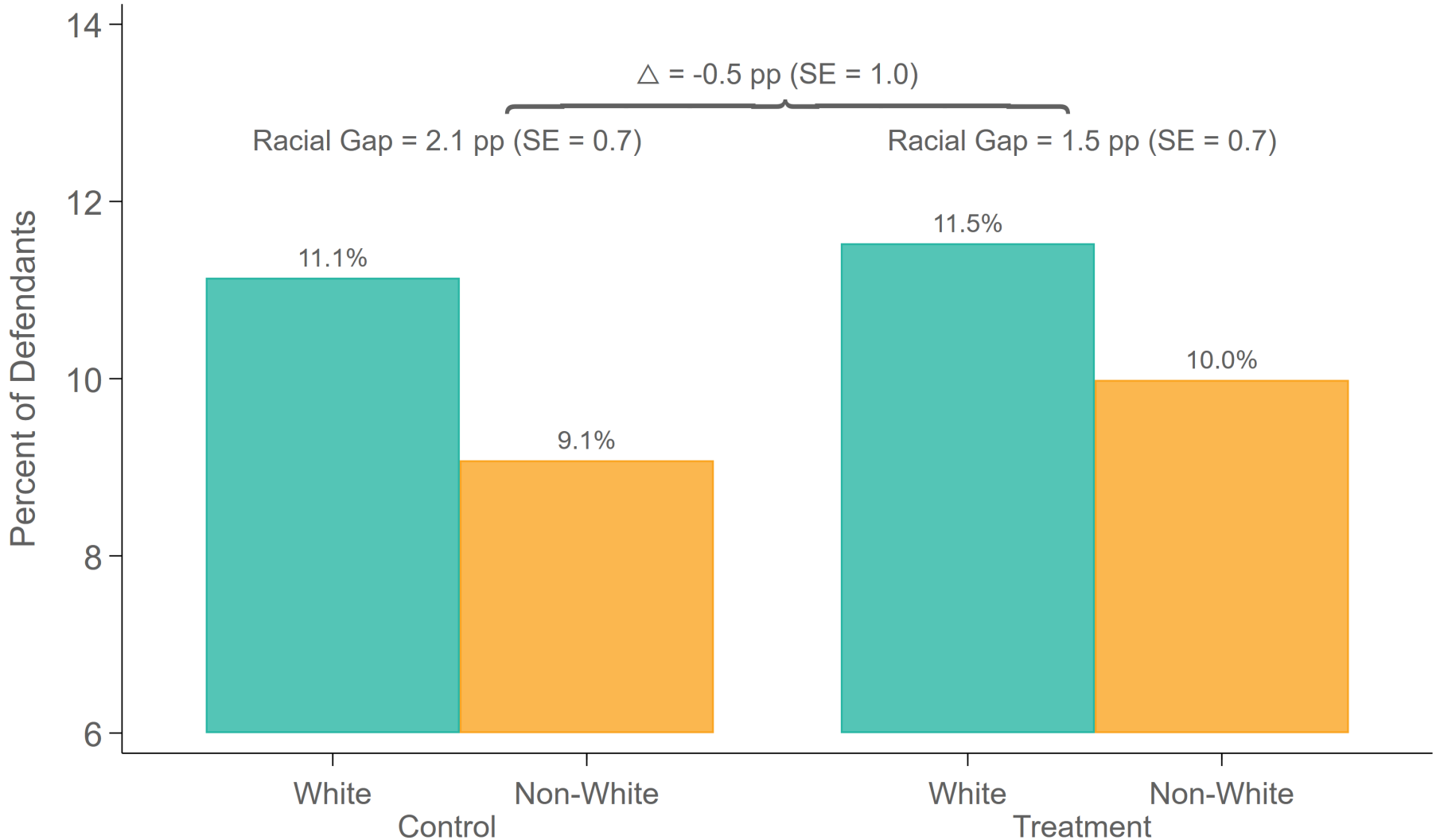
Note. This figure plots release on recognizance (ROR) by treatment status and defendant race (N=51 judges, 30k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White Misconduct Rates



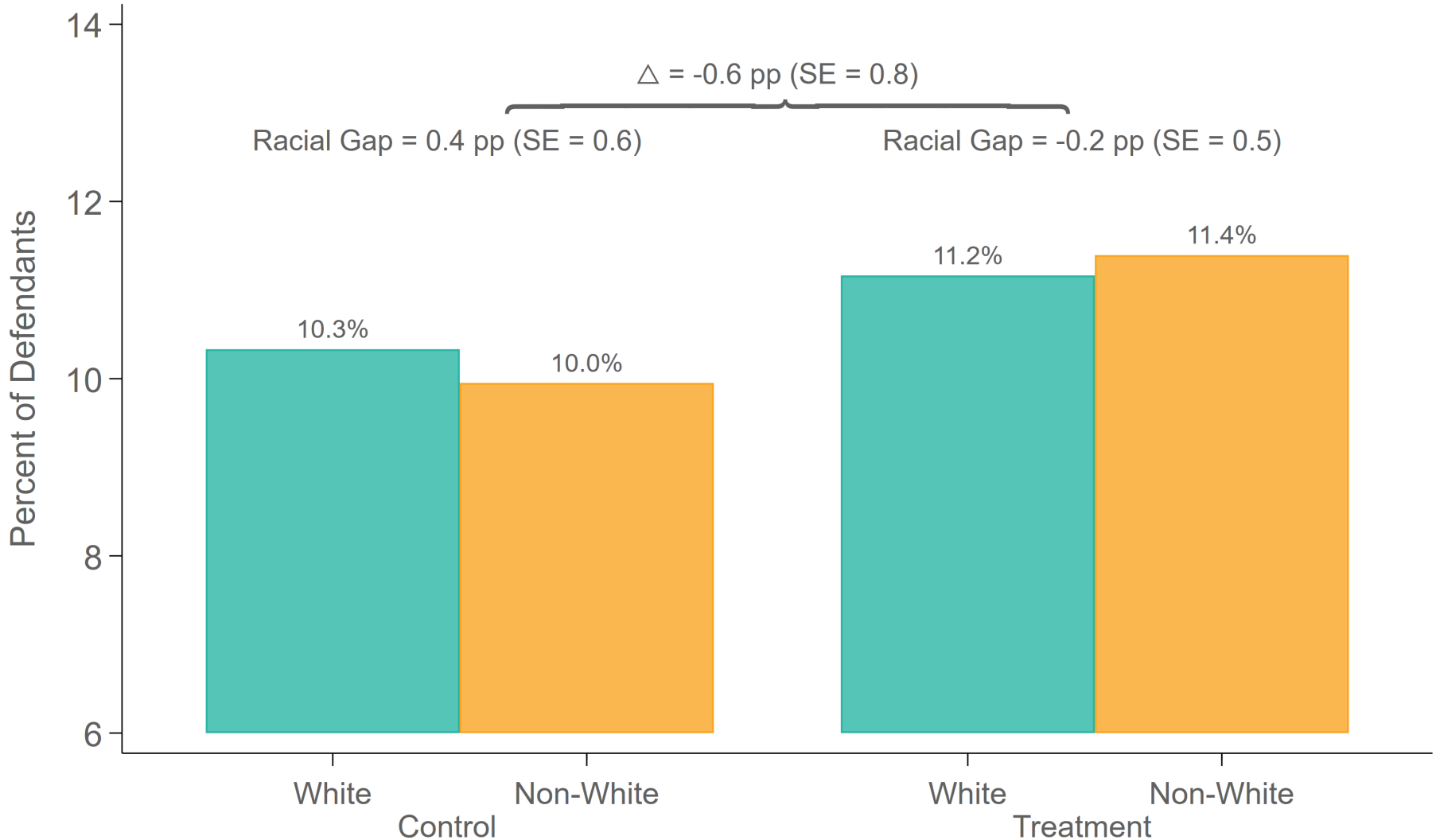
Note. This figure plots any misconduct (FTA or NCA) outcomes by treatment status and defendant race (N=51 judges, 25k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White FTA Rates



Note. This figure plots FTA outcomes by treatment status and defendant race (N=51 judges, 25k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

White vs. Non-White NCA Rates



Note. This figure plots NCA outcomes by treatment status and defendant race (N=51 judges, 25k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for court-by-time effects. Standard errors are clustered by judge.

NYC Intervention Materials

[redacted]

HARVARD LAW SCHOOL

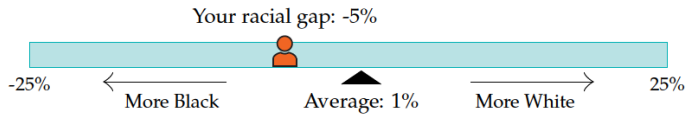
Your Bail Decisions vs. Other NYC Judges

This comparison is based on criminal arraignments in the Bronx Criminal Court from January 1, 2023 to March 31, 2023.

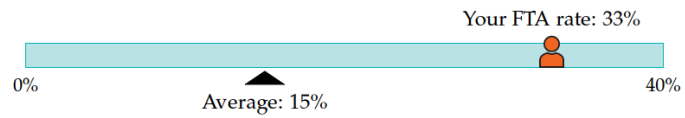
1. You release more defendants than the average judge.



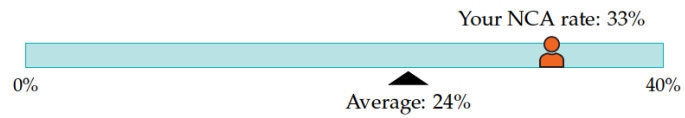
2. You release more Black defendants than White defendants.



3. You release defendants who are more likely to fail to appear (FTA) for court than those released by the average judge.



4. You release defendants who are more likely to engage in new criminal activity (NCA) than those released by the average judge.



Questions?

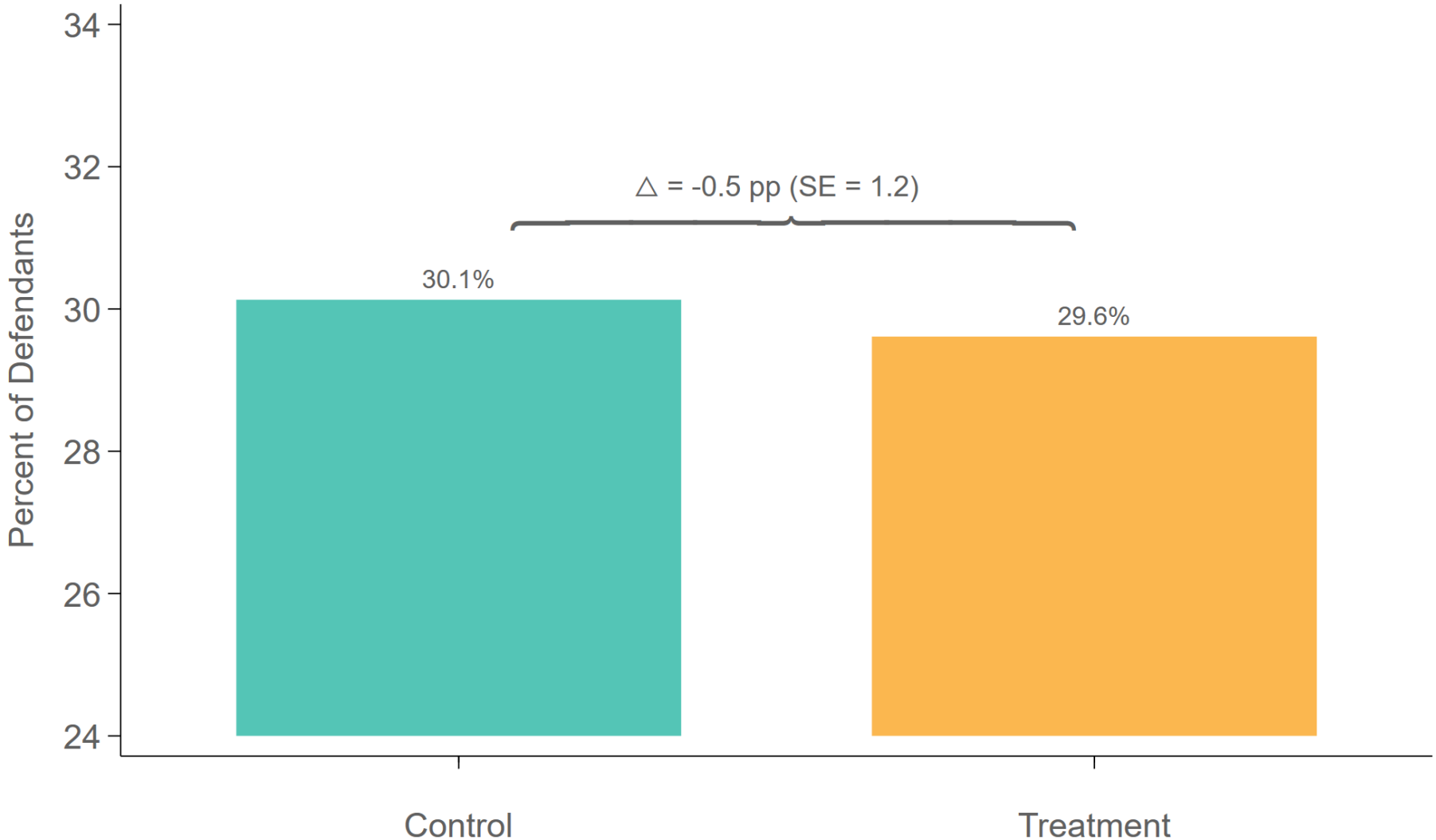
Contact us at 617-496-4477 or pretrialstudy@hks.harvard.edu.

NYC: Baseline Case Outcomes

	Control Mean	Treatment Mean	p-value
Release for All Defendants	92.3%	91.0%	0.093
Release for White Defendants	93.1%	92.6%	0.345
Release for Non-White Defendants	91.8%	90.0%	0.048
FTA for All Defendants	13.0%	13.8%	0.520
FTA for White Defendants	14.0%	13.5%	0.730
FTA for Non-White Defendants	12.5%	14.0%	0.238
NCA for All Defendants	21.8%	21.7%	0.908
NCA for White Defendants	21.3%	20.0%	0.183
NCA for Non-White Defendants	22.1%	22.8%	0.324
Cases	63,207	81,797	145,004

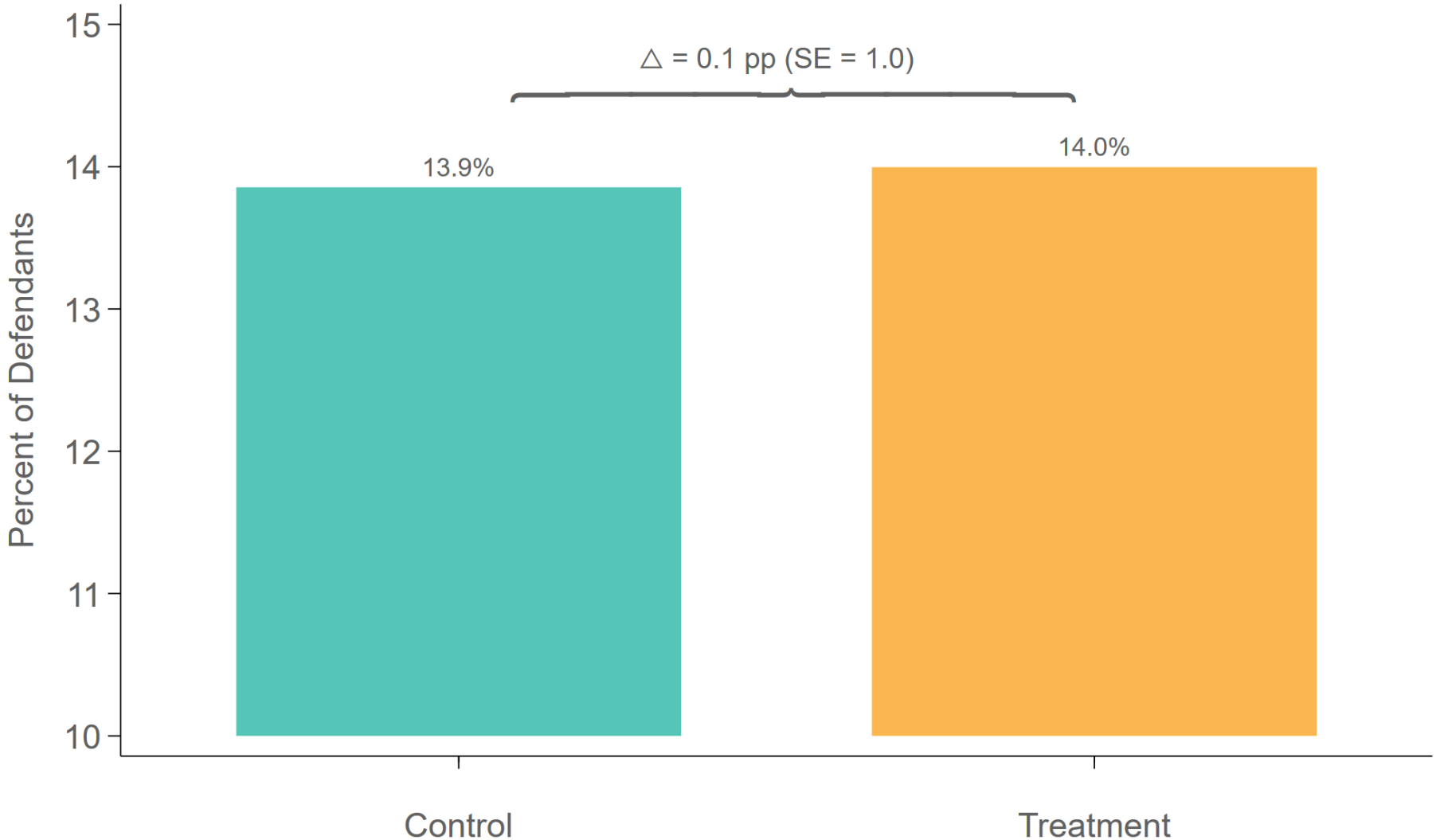
Note. This table reports baseline case outcomes for the 106 judges in our analysis sample.

Misconduct Rates in New York City



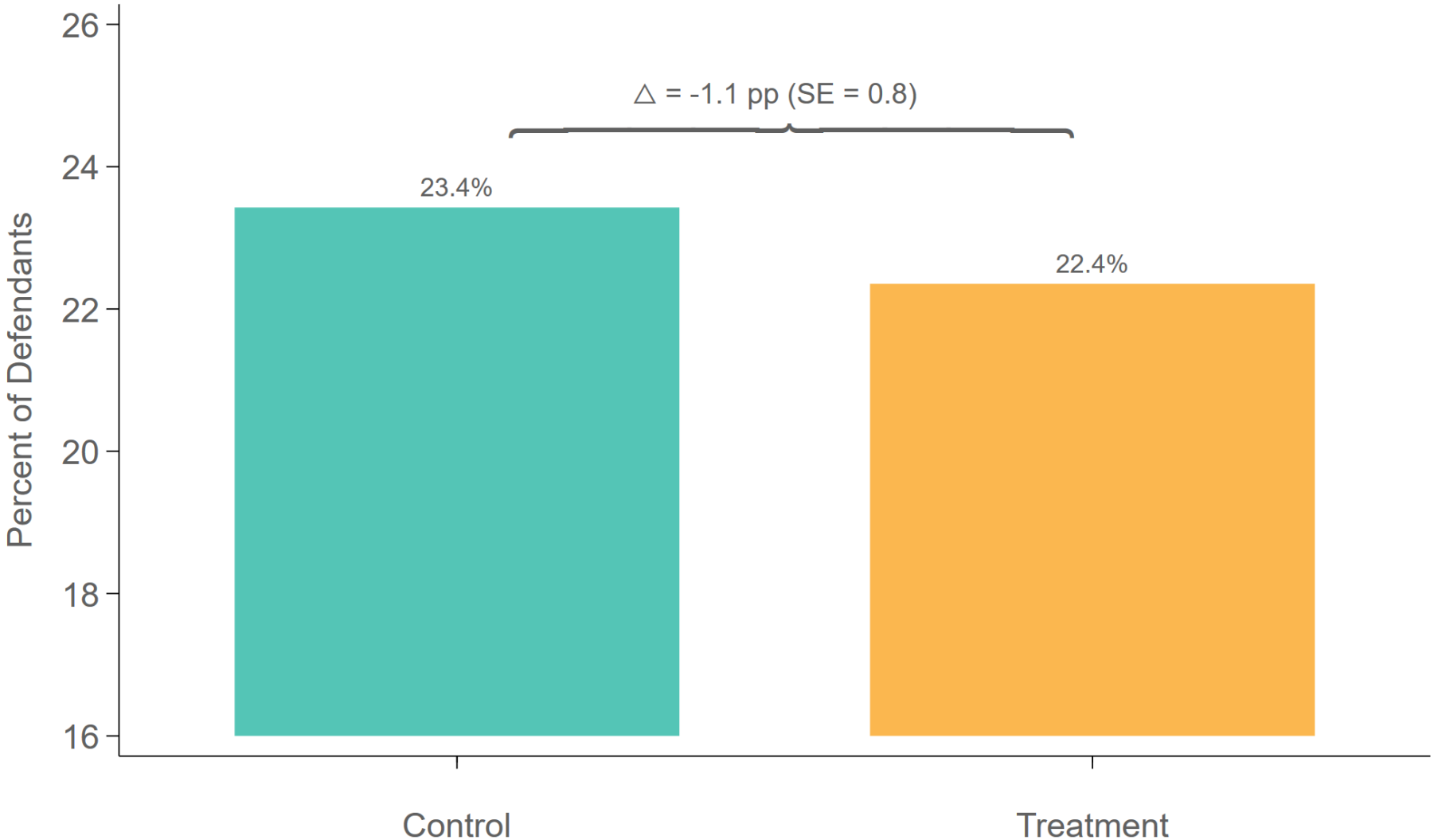
Note. This figure plots any misconduct (FTA or NCA) outcomes by treatment status (N=94 judges, 51k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for time effects. Standard errors are clustered by judge.

FTA Rates in New York City



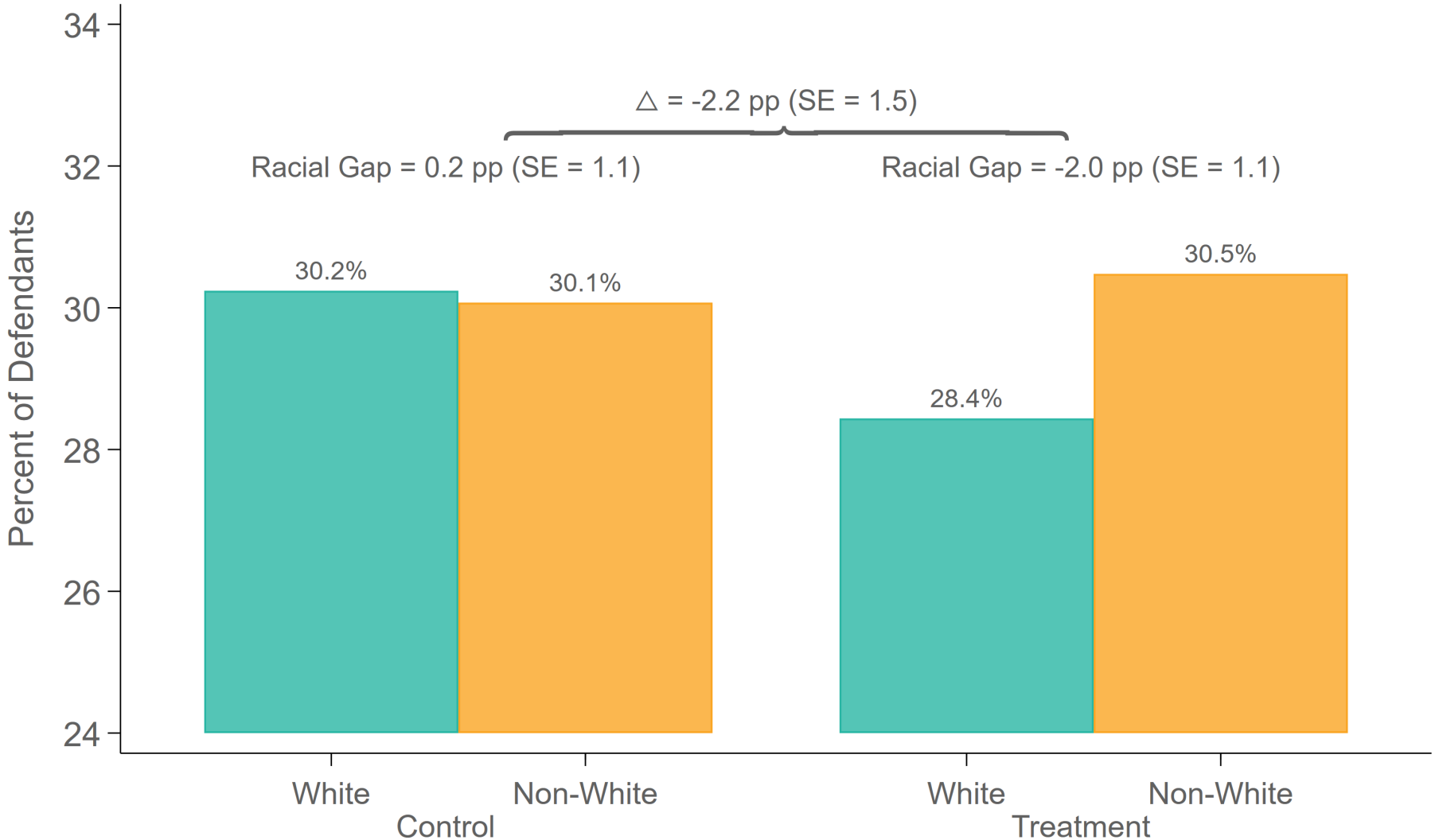
Note. This figure plots failure to appear (FTA) outcomes by treatment status (N=94 judges, 51k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for time effects. Standard errors are clustered by judge.

NCA Rates in New York City



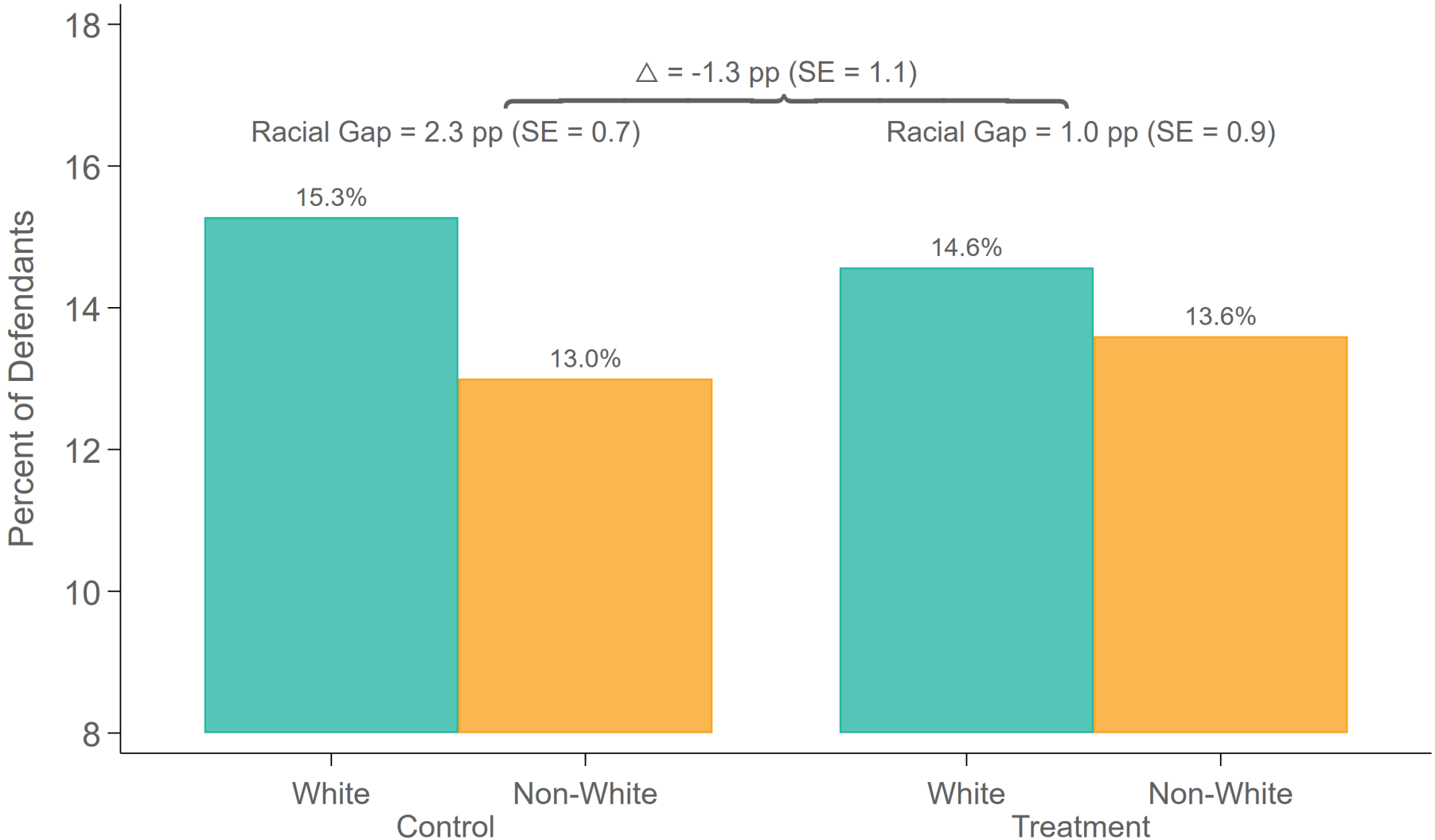
Note. This figure plots new criminal activity (NCA) outcomes by treatment status (N=94 judges, 51k cases). We report the control mean, with the treatment mean calculated from a regression of case-level outcomes on treatment status controlling for time effects. Standard errors are clustered by judge.

White vs. Non-White Misconduct in New York City



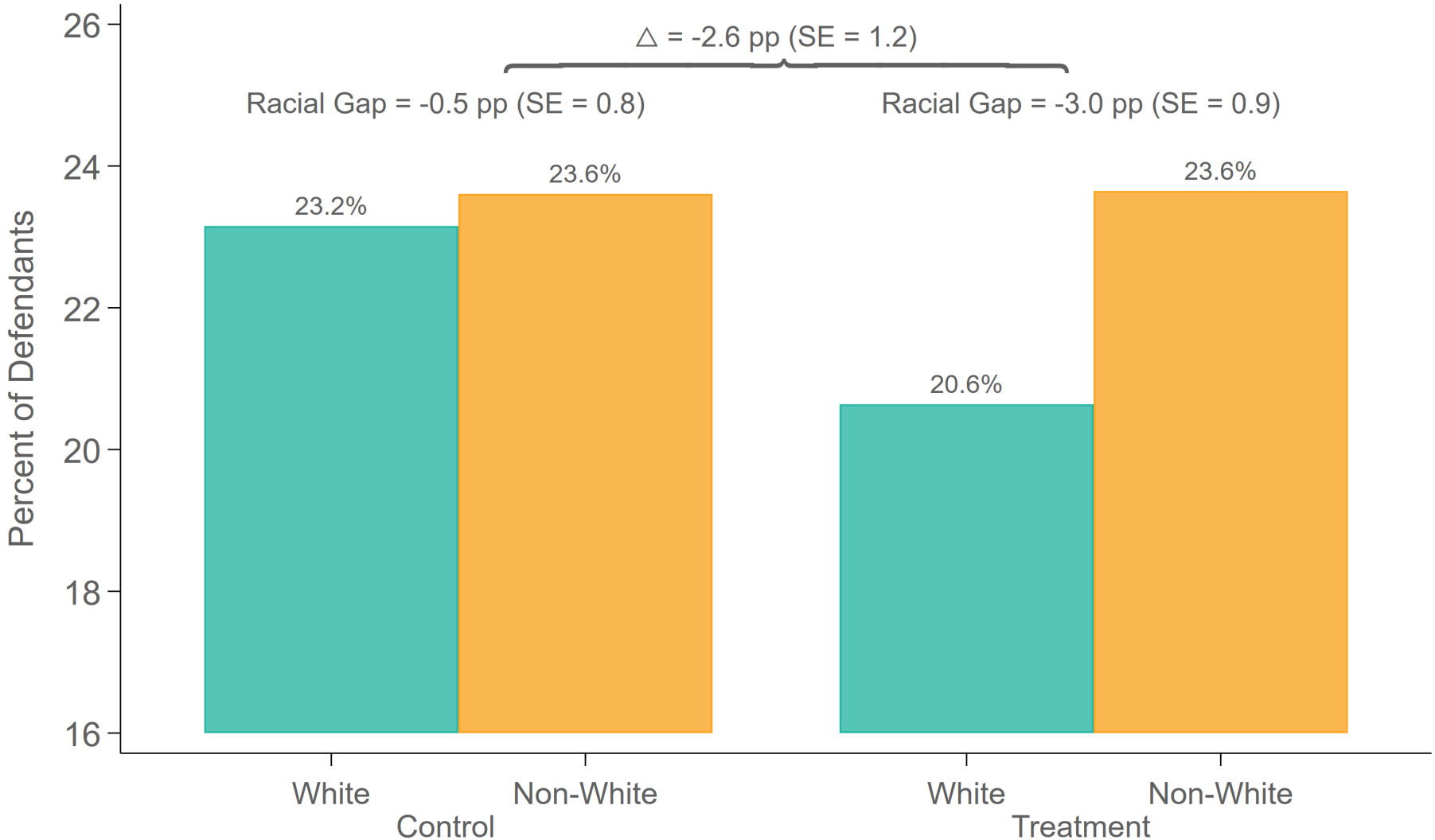
Note. This figure plots any misconduct (FTA or NCA) outcomes by treatment status and defendant race (N=94 judges, 51k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for time effects. Standard errors are clustered by judge.

White vs. Non-White FTA Rates in New York City



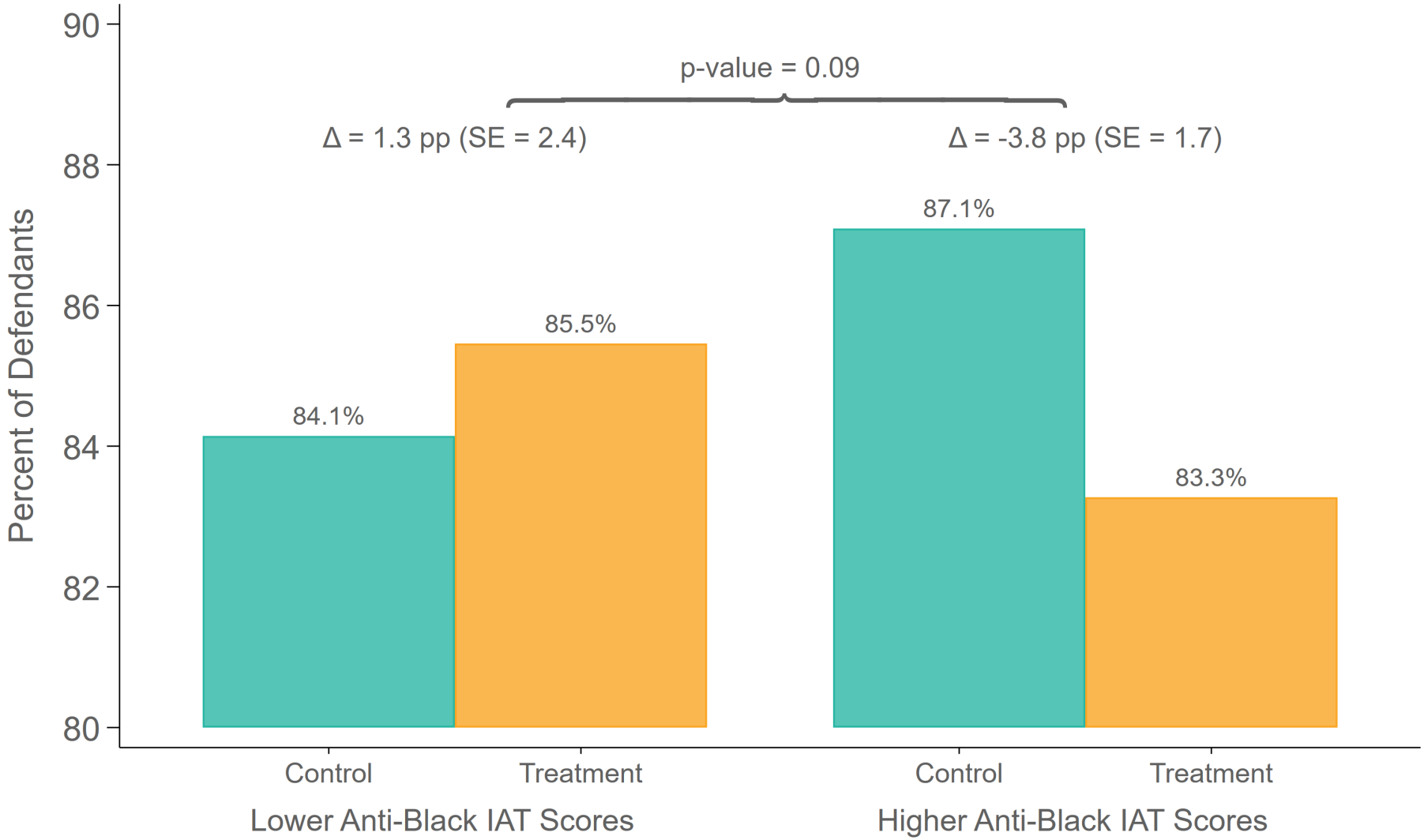
Note. This figure plots FTA outcomes by treatment status and defendant race (N=94 judges, 51k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for time effects. Standard errors are clustered by judge.

White vs. Non-White NCA Rates in New York City



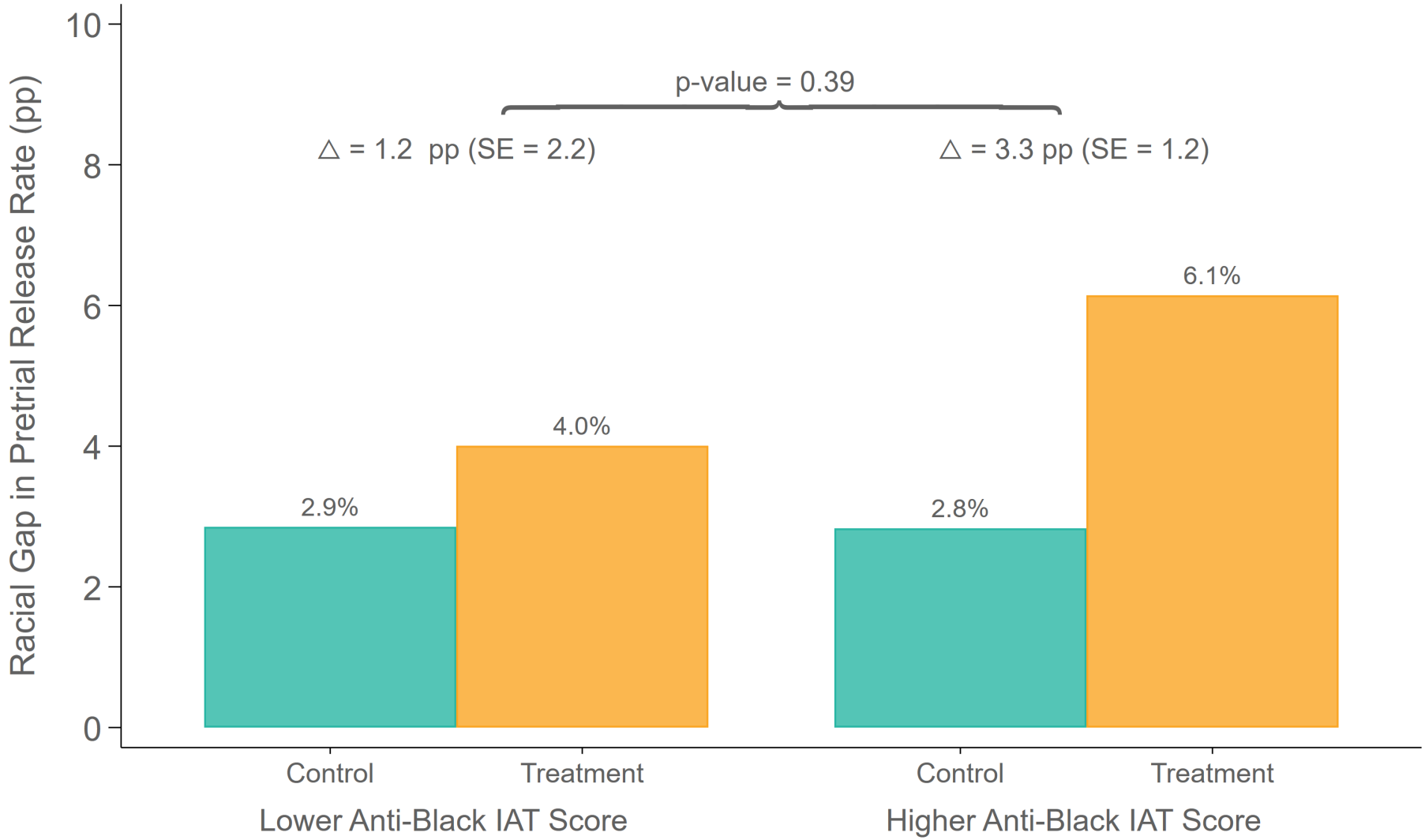
Note. This figure plots NCA outcomes by treatment status and defendant race (N=94 judges, 51k cases). We report the control mean for Non-White defendants, with all other estimates calculated from a regression of case-level outcomes on treatment status interacted with race, controlling for time effects. Standard errors are clustered by judge.

Release Rates by Baseline IAT Scores



Note. This figure plots release outcomes by treatment status and baseline IAT scores (below median vs. at or above median) (N=42 judges, 24k cases). We report the control mean for judges with lower anti-Black IAT scores, with all other estimates calculated from a regression of case-level release on treatment status interacted with an indicator for judge having a higher anti-Black IAT score, controlling for court-by-time effects. Standard errors are clustered by judge.

Release Rate Gaps by Baseline IAT Scores



Note. This figure plots the racial gap in pretrial release by treatment status and judge baseline IAT group (below median vs. at or above median) (N=42 judges, 24k cases). We report the control mean racial gap for judges with lower anti-Black IAT scores, with all other estimates from a regression of case-level release on treatment status interacted with race and an indicator for judge having a higher anti-Black IAT score, controlling for court-by-time effects. Standard errors are clustered by judge.