We thank the reviewers for the thorough and thoughtful reviews. Below are responses to the main comments.

(R2) Boston experiment: We agree with R2 that equalizing false positive rates could be helpful in some practical applications regarding criminality. We ran an experiment where we added the constraint of equalizing false positive rates in addition to equalizing true positive rates simultaneously. This constraint is also known as *equalized odds* (Hardt et al. 2016) [22] and has some precedent for use in the policing context.[1] Enforcing equalized odds additionally tests the proposed robust approaches on stricter fairness criteria compared to the "easier" criteria of equalizing true positive rates alone (another concern pointed out by R2). The model can no longer arbitrarily add positive predictions to increase the true positive rates, as this would also increase the false positive rates. Results are reported in Table 1 below.

Importantly, the label in the Boston dataset is whether an individual was searched/frisked in the past, and does not indicate whether the search was justified. If a predictive model trained with this label were used to determine whether to search/frisk someone in the future, then that application itself could harmfully carry forward biases in past search/frisk decisions in the training set. We want to emphasize that we do not endorse the application of the predictive model in this way. To ensure that we are not implicitly endorsing any problematic downstream applications of this model, we have included a paragraph specifically highlighting this label bias issue in the description of the Boston experiment.

(R2) Additional dataset: We ran an additional experiment with the NYPD Stop, Question, and Frisk dataset[2] (SQF), and train a classifier in the same way as prior work [1][3][4] to predict whether an individual when stopped does *not* possess an illegal weapon. Unlike the Boston dataset, this label is not subject to the same historical decision bias (though it still suffers from sampling bias as we have labels only for those who are searched). Table 1 contains results of enforcing an *equalized odds* constraint. As with the Boston experiment, the proxy race groups are estimated from the *precinct* feature in combination with public US census data, with a similar overall noise level 0.53 (see Appendix F.1.2). We'll be happy to include these results, either in addition to or as a replacement for the Boston experiments.

Table 1: Error rate and true positive rates (TPR) / false positive rates (FPR) constraint violations on test for Boston (top three rows) and SQF (bottom three rows) (mean and std. err. over 10 splits).

Algorithm	Unconstrained	G known	Naïve	DRO	Soft assign.
Error rate (Boston)	0.278 ± 0.001	0.290 ± 0.001	0.290 ± 0.001	0.315 ± 0.001	0.320 ± 0.001
TPR Max G viol.	0.059 ± 0.012	-0.010 ± 0.005	0.008 ± 0.005	-0.029 ± 0.003	-0.018 ± 0.002
FPR Max G viol.	0.007 ± 0.002	-0.011 ± 0.001	-0.007 ± 0.002	-0.008 ± 0.008	-0.020 ± 0.001
Error rate (SQF)	0.148 ± 0.004	0.181 ± 0.008	0.165 ± 0.004	0.343 ± 0.015	0.382 ± 0.033
TPR Max G viol.	0.004 ± 0.007	-0.014 ± 0.007	0.043 ± 0.009	0.023 ± 0.013	-0.006 ± 0.012
FPR Max G viol.	0.112 ± 0.039	-0.006 ± 0.028	0.087 ± 0.012	-0.008 ± 0.025	-0.020 ± 0.022

(R3) Upper bound on fairness violation: R3 raises the question on scenarios where the upper bound on the fairness violation w.r.t the true group using TV distance is tight (Theorem 1). In particular, Theorem 1 is tight for the family of functions that satisfy $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq 1$. This condition holds for any fairness metrics based on rates such as demographic parity, where h is simply some scaled combination of indicator functions. However, for a different particular given set of h, Theorem 1 may not be tight. We have clarified this discussion in the main text.

(R3) Cases where the noisy proxy variables $\hat{G}$ are high-dimensional compared to the true groups G: We agree with R3 that an evaluation of the SA approach when $\hat{G}$ and G have different dimensionalities would be valuable in a future empirical study. Theoretically, this can be handled by both the SA and DRO approaches. In particular, for DRO, Lemma 1 can be generalized to $TV(p_j, \hat{p}_i) \leq P(\hat{G} \neq i | G = j), j \in \mathcal{G}, i \in \hat{\mathcal{G}}$, and the true group distribution p_j can be bounded in a TV distance ball centered at $\hat{p}_i$. It is interesting future work to compare the robust approaches when the noisy proxy variables have different dimensionality from the true groups. We've added this discussion to the main text.

(R1 & R4) Estimating the bound on TV distance: Lemma 1 provides a practical way to estimate an upper bound on the TV distance between the data distributions under the true groups and noisy groups from an auxiliary dataset. As R1 points out, the looseness of the bound will lead to over-conservativeness of DRO approach. Furthermore, as R4 points out, if estimates from the auxiliary dataset are off, the robust approaches will also not be calibrated correctly. We agree that these are both important to note to practitioners. Developing methods to better calibrate the DRO neighborhood, and further study of the impact of distribution mismatch between the main dataset and the auxiliary dataset would be valuable future work. We have added more thorough discussions on these limitations and future work in the main text.

[1] Nathan Kallus, Angela Zhou. Residual Unfairness in Fair Machine Learning from Prejudiced Data. *ICML*, 2018.

[2] `http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page`

[3] Zafar et al. From Parity to Preference-based Notions of Fairness in Classification. *NeurIPS*, 2017.

[4] Goel et al. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy *Annals of Applied Statistics*, 2016.