We thank the reviewers for their thorough reviews. As suggested, we will move some lesser used notations in the main paper to the appendix and include a terminology table. Below are responses to the main comments.

Reviewer 1. User study: The oracle can be an expert, a group of experts, or an entire user population. Our framework can be applied by posing classifier comparisons directly via interpretable learning techniques [36, 10] or via A/B testing [39] (section 8, point 3). For example, for internet-based applications, one may perform A/B testing by deploying two classifiers A and B with two different sub-populations of users and use their level of engagement to decide which of the two classifiers is more preferred. For other applications, we may present to the user, visualizations of the predictive rates for two different classifiers, along with textual explanations, and have the user provide pairwise feedback. See e.g. Figure 3 in ref. [45] in the paper, or Figure 2 in ref. [A] below for intuitive ways to visualize rates.

Real-data experiments: We'll certainly move the experiments in Appendix E to the main text, given the additional page. We note that most baselines use some form of elicitation or prior knowledge. We wish to make two points in clarification: (a) even for this simple example, the results show that some kind of elicitation is usually preferable to no elicitation, even when measured by NDCG, (b) perhaps more importantly, since we have generated the oracle's metric randomly according to Definition 1, we may have picked a case that is too benign. It is not difficult to construct realistic settings where a default metric orders the classifiers arbitrarily different at the top from the true oracle's metric.

Reviewer 2: We thank the reviewer for the positive feedback. We'll simplify the suggested notations. With regard to mismatch between oracle preferences and the metric, please see the response to Reviewer 6.

Reviewer 3. While we have dedicated sub-sections on the problem statement and background material in Section 2, we understand why the reviewer might have found it to be insufficient. To make our work more accessible to non-expert readers, we will use the additional page for (a) adding an end-to-end numerical example, (b) elaborating on the metric selection problem with the intended use-cases around Figure 1, and (c) giving more intuition about each elicitation step.

Reviewer 5. Thanks for the positive feedback. As suggested, we'll clarify the intended use-cases early on with more text around Figure 1 and include the additional reference as well. Thanks for the helpful suggestions.

Reviewer 6. Equivalence of classifiers and rates: Our goal is to elicit *group-fairness metrics* such as equal opportunity [14] that can be expressed as a function of group-specific rates of a classifier. Since the metrics we consider depend on a classifier only through its rates (lines 46, 116), comparing two classifiers on these metrics is equivalent to comparing their rates. The concern the reviewer raises is true in general with any group-based notion of fairness and is not specific to our setup. While there are alternate definitions of fairness that look at individual outcomes, and some early work on eliciting individual fairness metrics [18, 28], these are tangential to the current paper (see Section 7 for details).

Assumption of fixed groups: This is a standard assumption in the fairness literature (see e.g. [2, 4, 9, 14, 21, 24, 27]). We agree that eliciting the fairness metric when the groups change or are unknown is an important and practical question, but beyond the scope of this paper. Moreover, we can adapt the standard practice of treating each intersection as a separate subgroup to handle intersections of groups. Other ways of handling overlapping subgroups for eliciting group-fair metrics is a promising future direction. Thanks for the additional reference. We'll add a discussion on this.

Mismatch between preferences and metric. The class of metrics we handle is fairly broad and includes many common classification metrics such as weighted accuracy, popular group fairness metrics such as equal opportunity, and combinations of those. Indeed our approach is able to both recover existing metrics and discover new ones. Moreover, there is precedent in various application domains to modeling user preferences with a parametric form [e.g. B-D], including prior work on metric elicitation [16, 17]. We thus see our work as a significant first step towards eliciting more general classes of fairness metrics. On the question about feasibility, we allow all feasible rates, i.e. for which there exists a classifier; for elicitation, we exploit a subset of these rates (e.g. sphere $\mathcal{S}_\rho$) with specific feasibility properties.

Multiple user preferences. Our approach is robust to noise in the preference feedback (see Section 5). We can handle many common noise models including the one described in Definition 4, and a model where the feedback is only probably correct. One way to handle multiple conflicting preferences is to simply aggregate them into a single feedback (e.g. with a majority vote), and use the aggregated feedback for elicitation.

Reviewer 7. Local-linearity: The reviewer is correct that the metrics we consider are piece-wise linear functions of rates, but we needed to substantially extend prior work on linear metric elicitation to be able to exploit this structure. A key challenge was to select queries so that we can jointly elicit the performance and fairness violation by either zeroing out or linearizing these terms. Handling more general non-linear functions is a promising direction for future work.

Metrics: In addition to equalized odds, our approach can recover many common fairness metrics such as equal opportunity, error-rate balance, etc. (see lines 97–101), and can also be extended to handle demographic parity [2]. It would be interesting to explore metrics such as positive predictive value that are fractional-linear functions of rates [17].

Optimization w/o elicitation: Thanks for the interesting pointer. We'll add a note on this.

Added References: [A] "Visualization of Confusion Matrix for Non-Expert Users" [B] Abbasi-yadkori et al. "Improved Algorithms for Linear Stochastic Bandits" [C] Valkenhoef et al. "Entropy-optimal weight constraint elicitation with additive multi-attribute utility models" [D] Abeel et al. "Apprenticeship Learning via Inverse Reinforcement Learning"