We thank the reviewers for careful examination of our paper. Since there are no common concerns, we address individual concerns. For the rebuttal, we use references from the main paper plus some references added here.

Reviewer 1. **1. LFPM Elicitation and Significant Contributions:** In our experience, linear metrics are by far the most used in practice (see [A, 21] and references therein), so we chose to focus on this case. Even for the linear case, there are many subtle issues that we address – including a novel characterization of the space of confusion matrices, introducing and analysing restricted Bayes optimal classifiers, developing algorithms with theoretical guarantees, and showing robustness for the practical applications (noise analysis). We hope the reviewer agrees, in accord with the other reviewers, that these are significant contributions. We also note that the linear case is important for understanding more complex settings, though all the additional details are difficult to compress into eight pages. However, we agree with the reviewer that LFPM elicitation is still important, and therefore, instead of discarding it completely, we summarized it in Section 7 and discussed it thoroughly in the appendix to conclude our scientific contributions.

2. Assumption 3 and 4: These are sufficient conditions for DLFPMs (LFPMs) to be bounded and monotonically increasing (decreasing) in diagonal (off-diagonal) elements of the confusion matrices. This is detailed in proof of Proposition 5 (Proposition 7). It is equivalent to fixing $\|\mathbf{a}\|_1 = 1$, $a_i \geq 0$ for the diagonal linear case (Section 2.2). The only additional restriction for the linear-fractional case is $b_0 = \sum_i (a_i - b_i)\zeta_i$, instead of the derived condition $b_0 \geq \sum_i (a_i - b_i)\zeta_i$ (see line 614), which is sufficient to guarantee a unique metric bounded in $[0, 1]$ (instead of one of the equivalent alternatives). Note that most existing linear-fractional metrics satisfy these conditions [7, 11, 12].

3. Lower Bound: We conjecture that our bounds are tight (section 7), but we leave a proof for future work. Our initial analysis says that it requires an additional understanding of the query space. We hope the reviewer agrees that query complexity bounds are important even when lower bounds are yet unknown.

4. Factor of k: Notice that the error guarantee in Theorem 1 is in $\|\cdot\|_\infty$-norm; whereas, it is in $\|\cdot\|_2$-norm in Theorem 2. Thus, using standard norm bounds, it is clear that both have square root dependence on the number of unknown terms in $\|\cdot\|_2$-norm. We thank the reviewer for pointing this out and will clarify in the final version.

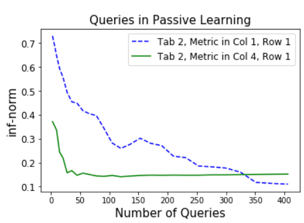

Figure 9: Passive Learning.

Reviewer 2. **1. Experiments:** Our experiments are primarily designed to empirically validate our theory. Since this is the first work on multiclass ME, we are unaware of any baselines. The suggested strategy of posing random queries is easily shown to require exponential time to achieve ϵ error (using ϵ-ball finite parcellation of the space of confusion matrices), thus is extremely query-inefficient. In Section 8, we outline several approaches which learn linear functions from pairwise comparisons in a passive manner [9, 6, 14] i.e. by first randomly collecting pairwise comparisons and then learning a linear function $\hat{\mathbf{a}}^T\mathbf{c}$. To verify the inferiority of the passive approach, we present the performance of [9] for the two metrics (for $k = 3, 4$) in row 1 of Table 2, and plot the error $\|\mathbf{a}^* - \hat{\mathbf{a}}\|_\infty$ in Figure 9. The plot is averaged over 5 random runs. We see that even after 400 queries the error is greater than 0.1 for the baseline; whereas, we only require 56 (resp. 84) queries for $k = 3$ (resp. $k = 4$) to achieve 0.01 error. While we chose not to compare to these trivial baselines, if the reviewer strongly feels these experiments are helpful for a broad audience, we are happy to add such experiments in the additional page of the final version.

2. Relevant Paper [B]: Comparison queries in [B] solve a different problem of actively finding a good classifier (wrt. the accuracy metric), compared to our problem of finding the oracle's metric. However, we believe some ideas from [B] may be relevant, and we would like to thank the reviewer for the reference. We will add it in the final version.

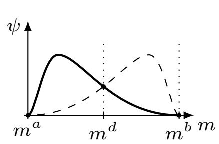

Figure 10: Two Queries

Reviewer 4. **1. Four Queries in Algorithm 1:** Unlike the standard binary search, we want to find the mode of a unimodal function using pairwise comparisons. Note that posing two queries in each iteration does not achieve the goal. As an example, compare the solid and dotted functions in Figure 10. Since the query responses will be same for both functions, we cannot decide the next search interval. Thus, we need more than two queries. On the other hand, we would like to thank the reviewer for pointing our unclear description of *unimodal*. Notice that due to Assumption 1, every supporting hyperplane of $\mathcal{D}_{k_1,k_2}$ supports a unique point on the boundary $\partial\mathcal{D}^+_{k_1,k_2}$ and vice-versa (Proposition 1); therefore, we indeed do not have flat regions. We will clarify this in the final version.

2. Difference in norms: The norms were chosen to best complement the underlying metric elicitation algorithm and vice-versa. For example, wlog, we can assume $\|\cdot\|_2$ normalization in Definition 1, but then the form of the solution becomes a little complex. If desired, we are happy to transform results to various norms using standard norm bounds.

3. ν, μ details: Thank you for the suggestion. We will add these details in the final version.

4. With high probability argument: When working with finite samples, we cannot guarantee that the estimate of confusion matrix $\hat{\mathbf{c}}$ will converge to the true $\mathbf{c}$ with probability 1 due to finite sample effects. Now notice that since the oracle response $\Omega(\hat{\mathbf{c}}, \hat{\mathbf{c}}') = \mathbb{1}[\phi(\hat{\mathbf{c}}) > \phi(\hat{\mathbf{c}}')]$ is a 1-Lipschitz function of the confusion matrices, we can guarantee correct feedback i.e. $\Omega(\mathbf{c}, \mathbf{c}') = \mathbb{1}[\phi(\mathbf{c}) > \phi(\mathbf{c}')]$ only with high probability (not with probability 1).

[A] Elkan, Charles. "The foundations of cost-sensitive learning." IJCAI, 2001.

[B] Kane, Daniel M., et al. "Active classification with comparison queries." FOCS, 2017.