Reviewer 1 *"What do...regret expressions represent? Why...max?"* The regret expressions represent how suboptimal agents are in each of the games (i.e. how much assumptions 1,4 in each of the games are violated). They are aggregated with a max since we would like to allow an ϵ-BNE in both $\mathcal{G}$ and $\mathcal{G}'$. Aggregating, for example, with a sum would mean that we allow for ϵ_1 and ϵ_2 equilibria in the games where $\epsilon_1 + \epsilon_2 \leq \epsilon$. We will make this clearer in the text.

"why is...revelation game the right relaxation?" We agree many relaxations are possible. We think the revelation game is attractive for the use-case of mechanism design since here the analyst knows the game structure clearly (they have coded the mechanism) but only has a model of the agents (possible utility functions, how well the agents optimize) and wants to be robust to the agent model being wrong.

"really only relaxing Assumptions 1 and 4... not sure that the ϵ-BNE...contain the equilibria implied by the correct reward functions" We agree with the reviewer's discussion of the relationship between RMAC and assumptions 1,2,3. We will clarify the text. Re: assumption 4, we agree that the fact that our procedure can be thought of as misspecification is not necessarily obvious, however, we have proven the following result: *Let $(\mathcal{G}, \mathcal{G}')$ be the real game/counterfactual game, let $(\mathcal{G}_m, \mathcal{G}'_m)$ be misspecified versions of these two games with same type/action spaces but $||u_m - u||_\infty \leq \frac{\epsilon}{2}$ (and same for u'). Let $\mathcal{D}$ be some data. If $r^* = (\hat{a}, \hat{\theta})$ is an equilibrium of the real revelation game corresponding to $(\mathcal{G}, \mathcal{G}', \mathcal{D})$ then r^* is also an ϵ-equilibrium of the misspecified revelation game corresponding to $(\mathcal{G}_m, \mathcal{G}'_m, \mathcal{D})$*. Note the converse is not true, so RMAC is a pessimistic estimate of relaxing assumption 4 appropriate for real world cases where we care about e.g. worst-case revenue. We will add this as a formal theorem to the paper.

"It would be useful to have more explicit discussion of the interpretation of ϵ. ϵ actually often has a natural scale since in many applications the units of the utility function have a natural scale. For example, in the case of auction where valuations are in dollars an $\epsilon = .5$ means individuals can behave in a suboptimal way that loses up to 50 cents relative to their optimum. We will add this example into the text.

"useful to see an example...[with] population of faulty agents..." The school choice example can actually be thought as doing this. The actions that we observe in the population can come from either a set of strategic agents or naive truthful agents. Depending on which underlying population generated the original data, we will get different counterfactual conclusions. The RMAC bounds show precisely this. We will expand the discussion to make this point explicitly.

"gap... between NP-hard exact solution..and RFP" Our results guarantee that if RFP converges then it converges to a local optimum. In some cases, this may not be the global optimum. Unfortunately, the MIP to compute the global optimum has $|\mathcal{D}|^2$ boolean variables so computing the exact solution is infeasible for instances beyond 10-20 data points. For such small cases a gap (or lack of gap) is unlikely to be representative of real world problems. One useful datapoint from our experiments is that across multiple initializations we found identical results in our auction/social choice experiments so it seems as though there are not multiple minima in these domains.

Reviewer 2 *"What does an element of $\mathcal{D}$ look like...single game, or multiple games?..types observed?'* d_i is an action played by an agent in the a single instance of the game. For example, if we consider the 'what would happen if we raised the reserve in the auction?' scenario then $\mathcal{D}$ each consists of an agent's bid when from one auction. In this paper we deal with 1 action per agent (though this is not required for the algorithm/theory). Importantly, types are **never** observed and must be inferred from data by making assumptions (in a standard model these are assumptions 1,2,4). We apologize this was unclear and will clarify the text.

"utilities to define the regret [not well defined]." We apologize and will fix this and other raised missing formalism. d_j is the action taken by the individual in $\mathcal{G}$. $\mathcal{D}_{-j}$ is the distribution of actions observed from everyone else. The regret for $\mathcal{G}$ of an individual with estimated type $\hat{\theta}_j$ is then the amount they lose by taking d_j instead of type $\hat{\theta}_j$'s optimal action (given that everyone else is behaving according to $\mathcal{D}_{-j}$). The regret for the counterfactual game is analagous except using the estimated counterfactual actions. We will make this clearer in the text.

Reviewer 3 *"not explained how this method differs practically from [1] and [2]"* We cite [1] in the text already and will add a citation to [2]. Both of these papers make their own versions of assumptions 1-4 - in other words, they can both be thought of as the solid arrow in Figure 4. By contrast, our goal is to relax these strong assumptions. As we discuss in the related work approaches like [1],[2] "allow for measures of statistical uncertainty..[but do] not allow analysts to check for robustness of conclusions to violations of assumptions."

We **do** make the comparison to the standard approach: we show the pure statistical uncertainty estimates (example: gray ribbons around ϵ=0 line in Figure 1). One can see that the statistical uncertainty in the counterfactual estimate is quite small relative to the non-robustness to small violations of the assumptions. We will make this clearer in the text.

"...it would be nice to know what new techniques/methods may be more broadly interesting." We provide 1) a MIP for small games, 2) a first-order method with some guarantees. Both of these are more broadly applicable than just our experiments and lead to many open theoretical questions.