We thank all the reviewers for their helpful comments and suggestions.

Reviewer 1 *RE: needing fair unlabeled data.* Just to clarify, when you say we need fair unlabeled data, do you mean unlabeled data (D) without selection bias? In that case, point taken, but we would argue that in some cases label bias a worse problem, and that to some extent, a good data scientist can help mitigate some of the known biases in the data D' to make it look more like what D. Moreover, if you could identify a subset of the data that is less biased, then maybe you could use this subset to improve the more biased subset through semi-supervision.

RE: synthetic data. Synthetic data is a powerful tool that is frequently employed in ML. For some scientific questions, like ours, it is the only empirical way to study them. If NeurIPS requires experiments on real world data then we would have to find another venue. Unfortunately, real-world data does not have the observable quantities we need to measure the true accuracy which we require to investigate the relationship between fairness and accuracy. We could follow R3's suggestion, but in the end, this still requires synthetic modifications to the data. We had initially considered a similar idea, but opted to discuss the perspective learning theory provides instead.

Reviewer 2 First, thank you very much for the thorough and thoughtful review, especially about the related work. It is very helpful and constructive.

RE: related work We actually agree with your characterization of the literature and see it as complementary to ours. Perhaps our language was sometimes too strong in our attempt to highlight what we perceive as a problem. E.g., the use of "most" instead of "many" may be severe. At the very least, "many" seems appropriate, and as you say, many fairness papers — including some highly cited papers by highly respected authors — fail to clearly state the assumptions. Take for example, Zemel et al *Learning Fair Representations*, one of the most highly cited papers in the area. The authors defer to two previous papers for accuracy reporting (Kamishima et al 2011 and Kamiran&Calders 2009), both of which discuss accuracy and fairness, but neither of which acknowledge the problem of label bias in that discussion. Conclusions are then drawn, discussed, disseminated (and repeated) without the assumptions needed to interpret them.[1]

It's true that some papers do mention label bias, and we directly quote the relevant passage from such a paper in the related work. Though, our perception is that more papers (than not) fail to mention label bias (and this might be where our characterizations of the literature differ). We believe that label bias is omitted frequently enough that it affects people's thinking on matters of fairness. For example, sometimes when presenting the semi-supervision idea to friends in the field for the first time, it is initially dismissed because of the alleged tradeoff. Once assumptions are accurately communicated and agreed upon, everything becomes copacetic again and the idea is accepted as realistic. To speculate: part of the problem is that in machine learning, we're so used to our gold standard labels being the indisputable truth (modulo Cohen's kappa), that it's easy to overlook label bias in fair ML, for which that's no longer true.

Thanks for pointing out our error about the Fish et al paper, which indeed does use simulated data to get access to the unbiased labels. They are exploring different questions than us, and so they don't end up investigating the accuracy fairness tradeoff, which we see as one of the main contributions of our paper and a significant distinguishing characteristic of our work.

RE: contribution We agree that adding yet another term to an objective function is not usually particularly novel, but in this case it first requires getting over the intellectual hump formed by the fairness accuracy trade-off. Maybe this is obvious, but the idea was initially met with criticism because of this purported obstacle. The synthetic bias is a mechanism that we use to investigate the fairness-accuracy tradeoff and we see the investigation itself as a key part of the contribution.

RE: other comments. (a) Agree that fair ERM might be a better term (b) Data set size was chosen to accommodate dimensionality while being small for computational expedience (we repeat each experiment 10x so there is actually 10 times more data in total), but in retrospect we could've made it larger (c) it's non-trivial (but possible via more sophisticated rejection sampling methods) to adjust for perfect class balance while turning the various other experimental knobs (like protected/unprotected ratio) since there are some dependencies between the experimental parameters; fortunately, the imbalance was not severe. (d) We will look at the language and try to simplify it for a broader audience.

Reviewer 3 This review very accurately identifies our specific contributions, and their relative importance to each other. We initially planned on doing something similar to your suggestion (3) in which we synthetically modified the COMPAS data, but decided to use the space for the learning theory example instead. Maybe a more compact way of conveying the same data would free up some room as you suggested in (2). With regard to (1), we cited that work as an example of in-processing and used the same objective function as our semi-supervised method, but on the testing set instead. Perhaps we should've cited a classification paper for in-processing too.

[1] To be clear, we have great respect for these works and these researchers, and are simply using them to illustrate the point.