We thank the reviewers for their thoughtful comments and suggestions and we respond below to some concrete questions/comments that were raised.

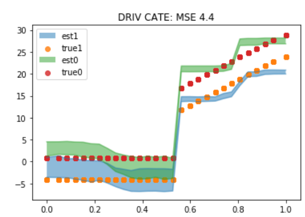
Response to Reviewer #1.

Sample splitting. We agree that it was our omission not to point out more clearly the relevant references on sample splitting. The most related is the sample splitting performed in prior papers that use the notion of Neyman orthogonality such as [1, 2]. We will add the relevant discussion and the relation to the papers recommended by the reviewer in the revision.

Response to Reviewer #2.

Width of confidence intervals. The intervals are large for two primary reasons: 1) the strength of the instrument, which is a typical source of high variance in IV regression, 2) the correlation among the features. Despite these sources, the number of samples was still large enough to identify several statistically significant non-zero coefficients.

Further semi-synthetic experiments. We have conducted several experiments with several functional forms for the HTE. We chose to depict a representative subset in the main text and the supplementary material, focusing primarily on the real world data. We will definitely augment with more semi-synthetic analysis in the revision and we have already included in our submission and will make public our code, that contains an easy to use Jupyter notebook, where one can play with the functional form of the HTE (see W_DGP_analysis.ipynb). For instance, we depict below one example of a DGP with a step-wise and discontinuous HTE and how our DRIV with a random forest final stage HTE model performs as a quick non-linear qualitative example (there are 9 features of which only 2 are relevant. One is binary and one continuous; the two dotted lines correspond to the HTE functions for the two values of the binary feature and the shaded lines correspond to the estimates recovered by DRIV).



Further benchmarks. We compared primarily to the orthogonal IV approach for ATE estimation of [2]. This is denoted as DMLATEIV. We found DeepIV to be unstable for our problem and we did not include the results. One major discrepancy is that the existing implementation of DeepIV uses mixture density networks to fit the distribution of the treatment conditional on the instrument. However this is an overkill when the treatment is binary (which is our primary case). In that case the advantage of DeepIV (which is primarily that it can capture non-linear relationships with respect to the treatment), is lost, since the outcome is linear in the treatment, without loss of generality (due to the binary nature). In such settings it suffices to just fit a regression that predicts the mean treatment conditional on the instrument and the features, as opposed to a distribution. This is essentially our DMLIV method; on top of that DMLIV also performs residualization which leads to extra robustness. For these reasons we omitted DeepIV experiments. However, we will definitely add such experiments in the revision.

Title. We will change to a more elaborate title, e.g. "Orthogonal Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments".

Response to Reviewer #3. We note that the main methodological novelty of the the DRIV method is to produce a Neyman orthogonal loss for the HTE estimation problem with instruments. The existence of such a loss was not known in the econometrics literature (for instance it was explicitly posed as an open question in [3]). The comment of [2] on the availability of Neyman orthogonal scores for ATE via the work of Robins and Rotnitzky is primarily for the conditional exogeneity setup and not for the IV setup with an unobserved confounder. Doubly robust estimates for the IV setup were developed in the work of [4]. However, this work assumes constant effect. We will improve upon our discussion on related work in our introduction and how our method contributes to the econometrics literature.

References

[1] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally Robust Semiparametric Estimation. *arXiv e-prints*, page arXiv:1608.00033, July 2016.

[2] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[3] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[4] Ryo Okui, Dylan S Small, Zhiqiang Tan, and James M Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205, 2012.