Review of "Estimating population average treatment effects from experiments with noncompliance" for *The Journal of Causal Inference*

I appreciated the opportunity to re-review this manuscript. I believe that the manuscript and contribution are much clearer. I have outline two remaining points of clarification I'd like to see below.

- **Definition of SATE:** The authors define SATE in their simulations as "the ITT effect estimated from the RCT sample adjusted by the sample compliance rate" (pg. 11 for example) This is contrary to how we'd usually define SATE and the complier average treatment effect (whatever acronym would be most appropriate in this particular example). The usual SATE estimator would be:

  $$\mathbb{E}[Y_i 1 - Y_i 0 \mid S_i = 1]$$

  But the authors are estimating the sample CACE, which is:

  $$\mathbb{E}[Y_i 1 - Y_i 0 \mid S_i = 1, C_i = 1]$$

  I think that the authors are correct that the sample CACE as it is the sample analog to the population parameter they are studying, but it is very confusing to see it described as the SATE. Clarification would help, or perhaps a different acronym.

- **Estimation of PATT:** I am still somewhat confused about the PATT estimator the authors use–as I understand it, they are estimating Y1 and Y0 in the RCT (based on actual treatment received) and then projecting this on to the population units to received treatment. This seems like it would require additional assumptions above the ones outlined in most of the extant literature that rely on reweighting methods. Usually these methods leverage the exchangeability in the randomized trial before projecting (more or less relying on similar estimates in small subgroups or strata), but here the authors implicitly adding an ignorability assumption within the experiment in order to assume they've properly (non-parametrically) modeled Y1 and Y0 in the face of non-compliance. In the RCT, receipt of treatment is driven by the C_i equation, but in the population it is driven by the T_i equation (and maybe also C_i). It is unclear to me, then, if the PATT estimator is performing poorly because of failure of the ignorability assumption in the sample, modeling assumptions, or the confounding? I think making the additional necessary assumption clear would at least help clarify why this method is more sensitive.

Small points:

- Pg. 2: "one might simply weight the PATT estimate by the population compliance rate in order to yield a population average effect of treatment on treated compliers". I found this confusing, because isn't the PATT estimate already projected on to those who received treatment (i.e. the compliers). I am not sure this would work, but it isn't a more appropriate thought experiment to reweight/project the ITT effect to the whole population and then divide by population compliance?
- Pg. 5 "RCT study designs that apply restrictive exclusion criteria may increase the likelihood that there are unobserved differences..." This sentence was a bit unclear to me. I could could see that the strict inclusion/exclusion violate positivity. But why would strict inclusion/exclusion criteria necessarily make for unobservable differences? In medical trials where the inclusion/exclusion criteria are known , it seems that would be preferable to an unknown, even if weaker, set of criteria. I think that the authors intuition is correct we might be somewhat more concerned, but I'm not sure the point is so straightforward.
- Pg. 5: footnote 3 has "Assumption 3 and 3"
- Pg. 14: Typo: "as closely than"
- Pg. 16: I thought the sentence "An important question for policymakers is whether..." was nice motivation that could be emphasized in the intro.

- Pg. 18: The authors state that interference is less likely in this RCT because of HH level treatment–how is this addressed in the analysis, and in particular, what does it mean for the definition of the population?