Dear Uri,

Many thanks for your interest in our work, and this important topic. We read your comments with interest.

First, it is very helpful that you acknowledge in your post that you were indeed involved as one of the four expert referees on this paper. As such your thoughts on this, and our responses to them, were fully aired and evaluated by the editor and associate editor in the context of the review process.

On the substance of “slow” versus “fast” p-hacking, this is indeed an interesting talking point. If some methods, policy areas or topic genres are systematically more amenable to fast p-hacking then that might plausibly generate smoother p-curves, and this seems like a useful thing to think further about (you may not need the simulation exercise – the possibility of the smoother p-curve is a fairly obvious artefact of your assumptions). Three observations on your assumptions;

1. Our main findings are to document the extent of p-hacking across non-experimental methods. It is unclear to us how slow versus fast p-hacking could explain much more bunching for IV at 1.96 and much less bunching for RDD.
2. The assertion that RCT as a method is plausibly *more* amenable to fast p-hacking does not seem natural to us. If anything we would expect the reverse. Consider a researcher working with outcome data from a large social survey such as the IHDS, which contains a plethora of health, education and attitudinal variables any of which could form the basis for an interesting research project. It seems plausible to think that such a researcher is much less hamstrung in switching between unrelated dependent variables from among the candidates (try a health variable, then an academic outcome variable, etc..) in search of significance than an RCT researcher would be. The same for trying unrelated treatment variables. Among the reasons for thinking this would be, (a) the ***much*** greater prevalence of pre-registration of RCT studies, which while not perfect does plausibly inhibit the sort of “jump to a totally different treatment variable/research question if you don’t like the results of the first one” that you seem to have in mind and, (b) the large fixed costs to rerunning a field intervention, compared to simply pulling an unrelated variable from the survey dataset. A third might be that RCTs are typically executed by larger teams of researchers, so any such dubious practice would require more conspiracy than would a sole researcher working alone on a laptop. A fourth might be that RCTs almost always have external funders, such that even ignoring preregistration there is almost always going to be an ex ante statement of dependent and independent variable before data is generated, unlike the norm where a researcher is accessing previously collected data. And while there are opportunities for fast p-hacking in observational studies there are equally opportunities for slow p-hacking in the data analysis phase of an RCT study – deciding which responses to remove, individual characteristic controls to include etc..
3. Of course, your heuristic iterative model of p-hacking as iterative is only one of a variety of ways in which one could set up a model significance-seeking research. In fact it seems less well-suited to thinking about RCTs than research using administrative or other extant data – there are large fixed costs to field execution which make “having another go” cumbersome. In terms of your example, then, if we impose an assumption of a single round of fieldwork it would require that all the candidate nudges be tried at the same time, and the less favorable results then discarded. That seems to us a rather counterfactual to how RCTs actually work, and would certainly imply no meaningful constraint associated with preregistration.

While we don’t necessarily agree with the way in which you are characterizing the process of research under the alternative methods, it is true that we are unable definitively to rule out the revisionist implication of your model which is that p-curve spikes just to the right of significance thresholds may in fact imply *less* p-hacking not more. However, as you know, the paper takes multiple approaches and brings together a number of different pieces of evidence. With this in mind we would encourage interested readers to look at the paper as a whole rather than selectively. For example the application of the Andrews and Kasy (AER 2019) methodology, which does not rely on looking for spikes near to thresholds, delivers consistent results.

On the substance of robustness checks and placebo analyses, we disagree with some of your claims. First, we sought to collect only coefficients of interest and exclude robustness checks and placebo analyses. Following your comments during the first round of revisions, we only collected estimates for the preferred bandwidth for RDD and excluded specification checks such as controlling for third or higher-degree polynomials of the forcing variable. Doing this had no meaningful effect on our conclusions or point estimates but was an important suggestion.

Second, we tackled another issue you pointed out by restricting the sample to the first table with results for each article. In other words, we are now comparing the first table of main results for each method, i.e., same number of observations across methods. It turned out that restricting attention in this way *increased* the size of the estimates for IV (rather than decreasing it). This can be seen from Online Appendix Table A22 for the estimates.

Third, in Appendix Table A32, we check whether our findings are robust to coding/data collection methods. We drop the tests for which the two independent coders could not easily reach agreement as to whether tests relate to main coefficients. In your blog, you mention that some of the tests in the paper on violence in Mexico (doi:10.1257/aer.20121637) should have been dropped since they are placebos. Note that these tests were flagged in our data set (see variable drop==1). We ended up including those tests in the main analyses but drop them in Online Appendix Table A32. One of the coders thought that they should be excluded while the other coder thought that the tests for the lame duck analyses could have been included since it is plausible that their could be an anticipation effect. Looking back at the paper, we agree that those tests should probably have been better excluded, in other words the first coder should have prevailed. But our methodology and transparency in coding allows interested readers to replicate all our tables and figures without any tests that are ambiguous. Doing this has no meaningful effect on results and conclusions.

On the ‘Into the weeds’ section and the number of modes in a distribution. It is good that you say that the characterization you give here is not definitively correct because we agree that it is likely unknowable what distribution, or mixture of distributions, the published z statistics come from. This is the main reason we perform the curve fitting exercise with both different distributions and different methods, finding similar results. What is the greatest issue is applying a monomodal distribution to something that is possibly (at least) bimodal.

Our first answer to this would be to point to the results of recent work by DellaVigna and Linos (NBER 27594) where the z curve for results generated by a nudge unit – and therefore not expected to reach statistical significance – displays monomodality. Second, bimodality from two distributions depends very much on the mixing parameter (conditional on the assumption you make that the means are sufficiently different, such as 0 and 2, as in your example). Should the population of z statistics come predominantly (rather than half as in your example) from the same distribution, then unimodality could plausibly be preserved. We are then left with the question, why, between non-experimental methods, would we expect to find a sufficiently different mixture of placebos and true effects (as in the example) such that the monomodal assumption generates a significant difference in excess test statistics for one method and not another?

Overall, we agree with some of your points and they are well taken. None of the methodologies delivers a definitive proof of the conjectures explored in the paper. Taken as a whole we do believe that most readers will find the congruence in results across the four different methodologies rather convincing.

Many thanks again for your continued interest in our work and for reaching out again with these opinions. You have done some fascinating work in this area and hope to continue to learn from you.

Sincerely,

Abel Brodeur, Nikolai Cook and Anthony Heyes