Response memo: AJPS – bridging the gap.

1. Editors comments:

Referee 1 points out that 2SLS does not really make a linearity assumption; rather, it only assumes that the instrument is orthogonal to the disturbances.

* Yes true, but for all practical purposes 2SLS is implemented as two sets of OLS, which in effect implement linear relationships.

This person says the text needs to formalize and motivate the proposed methodology more carefully and to engage the econometric literature on nonparametric IV estimation.

* We implement and refer to the extensive econometrics literature on non-parametric IVs for formalization.
* We have rewritten the introduction to motivate the use of nonparametric IV estimation better…

Referee 1 points out that the monotonicity assumption drawn from the causal inference literature is violated in the Monte Carlo analyses and the empirical applications.

* This is not correct: Monotonicity assumption (assumption 5 in JASA paper): comes from Rubin potential outcomes framework.
* Di(1) >= Di (0) – inequality for at least one unit = strong monotonicity.
* Monotonicity is implied in designs where those assigned to the control group are prevented from receiving the treatment.
* The monotonicity assumption rules out defiers: individuals who take the treatment if not assigned and not take the treatment if assigned.
* Monotonicity has no explicit counterpart in the econometric formulation (structural equation) but is implicit in the use of an equation with constant parameters for the relationship between the instrument Z and the treatment D. (the constant parameter assumption is stronger than needed).
* Monotonicity in the econometric case is violated if the instrument has different effects (or shapes of effects) for different units…
* e.g. treatment assignment leads to treatment for compliers but non-treatment for defiers.
* In the continuous case this would mean that e.g. the relationship between x and z is increasing for some units but decreasing for other units. Which is not the case, even if we use a non-parametric first stage and this will lead to a non-monotonic relationship between x and z – this relationship however has the same shape for all units.
* e.g. in the Krosnick example – time of interview has the same expected effect for all units – e.g. your dislike of Reagan increases approaching the scandal date and then decreases again. The problem would be if for some individuals the approaching scandal would reduce your dislike for Reagan and or going further away from the scandal in time would increase your dislike in Reagan.
* Or higher settler mortality increases the quality of institutions in some countries and reduces the quality of institutions in other countries.

Referee 2 provides a detailed review making the general point that the proposed estimator needs to be motivated more strongly.

* This corresponds to R1 and we have completely rewritten the introduction to motivate the use of a nonparametric first stage better

On specific points, this person says the text needs to differentiate unbiasedness and consistency. Currently, they are treated synonymously.

* Thank you! done

And on terminology, OLS and 2SLS are estimation procedures, not models.

* Thank you! done

Referee 2 says that the discussion treats the connection between the endogenous explanatory variable and the instrument as a structural relationship. In fact, it is merely the projection of the former on the latter.

* Not sure what this is supposed to mean?

Referee 2 says that no convincing argument is given about why the nonparametric method outperforms 2SLS.

* Really? Ok, we tried to make this point clearer in the text

He/she argues that the Monte Carlo example is trivial and questions whether the results are generalizable.

* We added more cases e.g. monotonic but non-linear case, e.g. decreasing marginal effects as in log-linear relationships

Referee 3 says that the discussion does not adequately balance coverage of IV estimation across different disciplines, and that this is particularly important given the intended audience at the AJPS. On a related point, there is quite a bit of relevant literature that is not cited in the text.

* We present an example from sociology and discuss the extensive literature in econometrics on nonparametric IV estimation

Referee 3 complains about "uneven" presentation of, and support for, the various ideas and strategies introduced in the text.

* Not sure what this means…

[We have tried to address R3’s concerns about uneven presentation of the IV procedure in the following ways.

1. Following R3’s suggestion our introduction now includes the conttibutions by Bollen and Dunning. Especially the inclusion of Bollen is particularly important because it offers a more comprehensive overview of IVs than the studies we have been referring to in the original MS. We thus refer to Bollen at various instances in the introduction and the IV overview section, as a key reference that highlights the usage of IVs in measurement and structural models.
2. We spend more time now on the introductory Figure, briefly discussing some of its key features.
3. Following R3’s advice, we now try to be more explicit about the fact that our choice of smoothers is far from comprehensive. Moreover, we try to justify why we opted for these three smoothers rather than others and we also acknowledge that as the MC evidence shows the difference between them is unlikely to be crucial in most empirical applications. What seems to matter, rather, is that any such smoother is used when a valid continuous instrument is available. ]

According to this person, more attention needs to be given to the desirable properties of instruments, including reliability and validity.

* We say that we assume validity, but now we are also showing results for when the instrument becomes slightly invalid
* I am not sure what reliability in the context of instruments means
* [As a way to address this concern we have now made more explicit the starting point of our discussion, which is that the instrument is valid. We explain from first section what we mean by valid and that our method does not (at least not directly) address the validity of the instrument, but rather concernates only on the efficiency, on how one can increase the strength of the first stage.]

**Reviewer #1:**

1. It is important to realize that the 2SLS is just a fitting procedure and does not describe the DGP. The IV estimate is based on the orthogonality of

instrument: E(Z \epsilon) = 0. From this, we solve \argmin\_{\beta} (Y - X \beta)^T Z Z^T (Y - X \beta), which gives you \hat\beta = (X^T Z (Z^T Z)^{-1} Z^T X)^{-1} X^T Z(Z^T Z)^{-1} Z^T Y, which turns out to be numerically equal to the 2SLS estimate. Note that there is no linearity assumption here. That is, we do not assume X is a linear function of Z. The only thing we assume is that E(Z \epsilon) = 0.

* Yes but in practice it is implemented as 2 OLS estimations and thus linearity is implied and all applied work is doing exactly this. We need to be clear about underlying assumption and practical implications.

Given this fact, the author needs to carefully motivate the proposed methodology. Are you assuming E(f(Z) \epsilon) = 0 where f is an unknown function? If so, how are you chosing f? That is, what is the objective function? The author would need to argue that f is chosen such that the statistical efficiency of the resulting 2SLS is maximized. To formalize this, write down the asymptotic variance of the IV estimator with Z\* = f(Z) instead of Z. Then, show that the choice of f the author advocates indeed minimizes this asymptotic variance.

* We are referring to the econometrics literature for this

2. Please engage with a large literature of nonparametric/semiparametric IV estimation in econometrics. I am certain that a similar procedure has been proposed by econometricians and it would be a good idea to clarify what's new and what is not.

* Yes. we are now citing and discussion the econometrics literature on nonparametric IV

3. From the causal inference perspective, which the author seems to adopt, the 2SLS estimation requires the monotonicity assumption (see the paper by Imbens and Angrist in JASA). This means that X must be a monotonic function of Z. It seems that both in Monte Carlo simulations and empirical applications, this assumption is violated.

* This is not correct: Monotonicity assumption (assumption 5 in JASA paper): comes from Rubin potential outcomes framework.
* Di(1) >= Di (0) – inequality for at least one unit = strong monotonicity.
* Monotonicity is implied in designs where those assigned to the control group are prevented from receiving the treatment.
* The monotonicity assumption rules out defiers: individuals who take the treatment if not assigned and not take the treatment if assigned.
* Monotonicity has no explicit counterpart in the econometric formulation (structural equation) but is implicit in the use of an equation with constant parameters for the relationship between the instrument Z and the treatment D. (the constant parameter assumption is stronger than needed).
* Monotonicity in the econometric case is violated if the instrument has different effects (or shapes of effects) for different units…
* e.g. treatment assignment leads to treatment for compliers but non-treatment for defiers.
* In the continuous case this would mean that e.g. the relationship between x and z is increasing for some units but decreasing for other units. Which is not the case, even if we use a non-parametric first stage and this will lead to a non-monotonic relationship between x and z – this relationship however has the same shape for all units.
* e.g. in the Krosnick example – time of interview has the same expected effect for all units – e.g. your dislike of Reagan increases approaching the scandal date and then decreases again. The problem would be if for some individuals the approaching scandal would reduce your dislike for Reagan and or going further away from the scandal in time would increase your dislike in Reagan.
* Or higher settler mortality increases the quality of institutions in some countries and reduces the quality of institutions in other countries.

4. A minor comment. I suggest that the author uses the coverage of confidence intervals rather than the overconfidence measure. The former is a standard way to evaluate the standard error of asymptotically unbiased estimates.

* Okay. Thank you for this comment, we have now changed to coverage instead of the overconfidence measure. The results remain unchanged.

**Reviewer #3:**

The conceptual backdrop does a decent job of establishing the foundations of the manuscript. However, it does not at present adequately balance the competing demands of providing necessary and sufficient detail. Although a core group of seminal works for IV estimation are covered from the econometric side (Angrist and co-authors, Bound et al., Wooldridge), the same level of detail is not shown for contributions from other social sciences. Overall, the presentation comes across as incredibly uneven. Citations are glaring absent from this section—it's an excellent opportunity for the authors to become a central piece cited in political science and thus a missed opportunity.

* [We are extremely grateful to R3 for this point. As a way to address it, we have extended our bibliographic coverage of IV estimation, including both Bollen’s work and Thad Dunning’s book on natural experiments. We cite both studies in the introduction and especially with respect to Bollen’s piece, we refer to it as a key source for readers interested in how IVs apply in measurement and structural models, thus when the movitation is not to address the endogeneity between *X* and *Y.*  See footnote 1 and page 4 of the revised manuscript. Although we would like to add more references, space limitations prevent us from doing so. Yet, the discussion about the Bollen article is particularly useful because it indicates interested readers into a very comprehensive study that discusses IVs in a different way than typically done in the econometric literature (i.e. without focusing only on the need to causally identify an effect of on Y).
* By the same token, we have extended the literature on non-parametric techniques, adding two important textbooks on the topic (page 11)]

* Thanks we have done a thorough search of work developing and applying IV estimation and have cited and discussed in the ms.

Let's take Figure 1 for instance. Indeed there has been phenomenal expansion in the number of studies using IV estimation in recent decades. The author(s) does not actually describe the figure nor does he provide enough detail to know whether all articles included in the count actually employ an IV approach, know the type of IV they have in their model, describe endogeneity, and so on. For Figure 1, the authors need to add detail to answer three questions: 1) why has there been this growth, 2) why the disparity between political science & sociology relative to econ, and 3) how do the studies reviewed differ with regard to their reliance on observational versus experimental data? Having reviewed a subset of literature for IVs every now and again, there are substantial differences across these studies regarding why and how they use IVs. Proportionally, I conjecture that at least ¾ of the studies use observation data, with a greater percentage in the political science and sociology subset of articles.

* Thank you for this suggestion, we have done a more thorough search of IV literature and now present developing IV methods, applying and what IV estimator is applied in the figure.
* [We are very thankful for this suggestion. In the revised introduction, our discussion of the Figure has been extended, trying to answer all three questions posed by R3: why is there still a gap between economics and political science/sociology? What explains the ascending trend? And what proportion of this work is based on experimental vs observational data? We now give more space to our description of the key features of the Figure, which we think has helped to better place our question in the existing literature.]

The authors need to greatly expand the intro so it does not read like it is tacked on to a paper that was in a previous life a research note rejected from a methods journal and submitted to the AJPS. As they proceed with revisions, the authors must make modifications to other sections accordingly.

* Thank you. We have fully changed intro to reflect development of IV and contribution. And we have adapted the following sections to this introduction.
* [Following R3’s suggestion the introduction has been expanded by half a page as a way to better motivate the key ideas of the paper. We believe that as a result of this change the paper now does not read as a research note.]

I am astonished that the authors do not discuss the review of instrumental variables published by Ken Bollen in 2012. His piece provides a cogent overview and review of instrumental variables including how they are treated via different approaches from simultaneous equations (as the authors do in this piece although they do not use the same terminology), structural equation models (not to be confused with the previously mentioned approach which is the one utilized in this particular manuscript though it is labelled structural equations approach), and limited dependent variable approaches like GMM-IV and Two-stage conditional probit. Surely the Bollen article made it into the count that is displayed in the figure. But, not discussing it is quite an oversight. A second glaring oversight given the reliance on Sovey and Green (2011) is the 2008 work by Thad Dunning—this is especially notable as it has an impressive theoretical model for which an IV approach is adopted.

* Thank you for the great suggestions, we indeed did not cite all the relevant literature
* Adapt discussion to terminology in JASA paper
* [We are thankful to R3 for this important line of criticism. Being ourselves political scientists, we had unconsciously prioritized the contributions in this discipline at the expense of other contributions in social sciences. In the revised version we refer to both Bollen and Dunning and we generally avoid discussing only references in political science or economics. Especially with respect to Bollen, we use this study as a very valuable source that discuss IVs not only from a causal identification perspective but also from the perspective of structural equation models and measurement models. Accordingly, following R3’s advice we have changed our terminology, embarking on Bollen’s typology and thus referring to our discussion of the IVs as a discussion that is motivated by the simultaneous equation framework. This is evident on pages 3, 4 and 5 of the revised manuscript.]

The unevenness in presentation extends also to sections where details and citations are sorely lacking and, in some instances, contradictory. Here are some examples: on page 10, the authors rely on a single work for establishing the utility of KRLS—how much has that single 2013 work been referenced? Much the same is the situation with kernel smoothing although the reference is from the late 1970s. And the other smoothing techniques section on page 11 should be expanded. Is the basis of the selection anecdotal or research-based? The current text does not lay this out linearly. In other words, the selection of these three techniques as the organizing principles of this piece is not particularly well established by a thorough review of alternate possibilities. I expect this to be the case for a manuscript to be published in a journal of this stature.

* [Following R3’s suggestion we have revisited section two in two ways:
  + We added a footnote already early in the section to advert the reader that our discussion here does not intend to encapsulate all non-parametric smoothers (footnote 5). Rather we briefly describe and later use few of them, which we choose because they are the first ones one comes across in any introductory textbook for non-parametric smoothers.
  + In Section 2.4 we explicitly motivate the choice by referring to the fact that kernel regression and local linear regression are the simplest and most intuitive ways of thiking about the procedure. We bring insights from Hardle (see footnote 10 accompanying this section) to further elucidate the idea that it is better to use a simple method for purposes of exposition. We also explain that at least asymptotically all other methods can be seen as equivalent to the kernel method and thus describing how kernel smoothing works helps to also provide a more general idea about how non-parametric smoothers work in general.]
* All these changes are now evident on pages 8 and 11 of the revised manuscript.
* Elias: we need to be more structured in selecting non-parametric first stages, perhaps we can rely on the Horowitz papers to establish which kind of non-linear first stage is best? Otherwise, we need to discuss the class of nonparametric estimators and potentially run MCs for all of them?
* Do we need to include GMM, GAM, and polynomial regressions?

In this iteration, the manuscript suffers from a number of organizational issues related to moving text from being in the main text to footnotes and appendices. The section on overfitting, for instance, needs to be reduced and made into an appendix. The text can include a single paragraph—the one that begins with 'To sum up, …'.

* Ok we have removed the section on over-fitting into an appendix
* Elias – if we get different reviewers we need to be very careful about this, the most questions I received was on over-fitting and optimizing bandwidth as well as cross-validation for the first stage (I guess we do not need to do this ourselves but cite the relevant literature on bandwidth optimization and cross-validation). I think over-fitting is important because in the limit the nonparametric first stage will capture completely the movement of the endogenous x – so over-fitting might induce bias.
* [In response to R3’s suggestion we have removed the text and the associated tables in the Appendix. Because, however, we believe that over-fitting can pose serious threats to inference with non-parametric methods, we briefly describe the conclusions of our MC analysis and our suggestion about best practices in this issue.]

Section 3 on statistical properties needs to be revised in accordance with the first section and the evaluative text that is sprinkled throughout the empirical applications in this version of the ms.

I am struck by the unevenness of the ms regarding the presentation and justification of the use of IVs regarding their basics. That is, an instrument must meet both theoretical and empirical criteria. The empirical bases are more developed than the conceptual basis, which is typically the case, but even that is uneven in how it is described and cited. The empirical tests noted on page 14 include typos—Durbin (not Dubin) and Hansen (not Hansan). In the same section, the authors note that they refrain from using the available tests in their analyses without providing justification. Dismissing the criteria for establishing the utility of an IV is silly as currently written in the text especially because they then contradict themselves a few pages later regarding the validity of the instrument (on page 19). By much the same token, the theoretical relevance of a potential instrumental variable is important since, as the authors themselves note, they are difficult and often viewed as impossible to find. More explicitly, the authors need to describe reliability and validity of the IVs in the empirical models, something largely lacking in this version but much needed in this area of scholarship generally. Simply stated, the 2 assumptions made in estimation relying on IVs—the first regarding correlation between IVs and the disturbance term and the second regarding correlation between the IVs and the troublesome regressor are about validity and reliability—both need to be examined using empirical checks. This point is not made clear in the ms and should be highlighted more.

* [We are grateful to R3 for this point. We have tried to address it by explicitly referring to the problem of instrument validity from the very beginning of the manuscript, page 3. There, we already try to motivate the key idea, namely that since no test is readily available for the validity of instruments it is understandable that researchers think about IVs having mainly the criterion of validity in mind. In this discussion we make it clear that our approach to the IV estimation is motivated by the need for causal inference, which is clearly the dominant perspective in the IV literature. Priorities might be different if for example IVs are used in a factor analytic model, as discussed by Bollen in his 2012 article (see footnote 2, page 3 in the revised manuscript). In this same discussion, we now try to clarify that our contribution is not to directly help researchers with a method to test or increase validity. Rather, our contribution is about how, given an instrument chosen on validity grounds, you can implement our estimation strategy as a way of increasing efficiency in the first stage.
* Although we do not use R3’s terminology (reliability), we clearly distinguish the two criteria referring to the former as validity (“correlation between IVs and the disturbance term”) and the latter as first stage (“correlation between the IVs and the troublesome regressor”). We stick to the terminology used by most textbooks but the idea is exactly the same.
* We are very grateful for spotting these typos, which we have hopefully now corrected, together with other typos we spotted during the revisions.
* We also thank R3 for spotting an apparent inconsistency between sections 3 and 4. We had not intention whatsoever to dismiss validity as a criterion. Rather, we think it is hugely important, perhaps the most crucial criterion for IV identification. What we refrain from placing much emphasis on is the over-identifying restrictions tests, which allegedly test the validity of the instruments. We explain that we do not want to put a lot of emphasis on these tests because they need more than one instrument and assume that one of these instruments is already valid. At this point we need to acknowledge that our own reading and motivation comes from IVs as tools for causal inference, or IVs as natural experiments in Dunning’s terminology (or as experimental IVs in Bollen’s terminology, although we believe that for estimation purposes whether the IV is based on observational or experimental data is indifferent). In this context, finding even one IV that fulfils the criterion of validity is a very hard task: finding more than one valid instruments is thus exceptionally difficult. This means that within the literature that motivates our contribution, such over-identification tests are not very useful. In the revised manuscript in Section 4 we try to make this distinction between validity and overidentifying restrictions tests much more clearly.]
* frequent as the
* Thank you for the comments. Validity of instuments cannot be tested that is the whole problem. That is why IV papers usually spend many pages on justifying the exogeneity of said instrument. There is no empirical valid test to establish validity of instruments. But we discuss tests now in more detail and explain why these tests cannot be used to test the validity of an instrument straightforwardly. In terms of what the reviewer calls reliability, I am not exactly sure what is meant by this but I assume it is the second condition of strength, or non-zero covariance with the endogenous RHS variable. This can of course be tested straightforwardly and we are discussing these tests now in more detail.

Additional suggestions for revision:

1) Why 3 examples? It seems possible to trim this to at most 2 in order to tackle the revisions noted in this review.

- [We appreciate the comment and have reduced the discussion of each example, and have relegated one example to an online appendix. We believe, however, that these three examples show the extent to which IV estimation is used in political science and economics. Each of the 3 examples allows us to point to different implications of the use of IV estimation: Example 1 illustrates how the method works with a very famous example from political economy. Example 2 highlights the problem with misspecification, when non-linearities are ignored in the first stage. Finally, example 3 showcases how the method can facilitate the researchers and lead them to simpler estimation strategies, thus dispensing the need for additional and often unnecessary assumptions.]

2) The number of figures needs to be reduced.

* [We are extremely thankful to R3 fot this suggestion. We have now reduced the number of graphs from 14 to 11.]

3) Table 1 needs to be revised—right now the reason for this set-up is unclear and it appears quite messy with decimals, organization of rows and columns without italics, etc..

* [Many thanks for this comment, we have tried to ensure all columns are aligned and we have added indications in the table and the note so that the reader can follow the table without further instructions in the text.]

4) The presentation of the figures in the text requires more careful consideration and possibly the inclusion of a table that sets the dimensions to be examined in a modified version of the classic 2 by 2 table that includes the relation between endogeneity, the instrument, and the estimator. This applies to Figures 2-9.

* [Many thanks for this. We have reorganized the figures to make them more self explanatory. Following R3’s suggestion we also introduced a table (Table 1) that guides the presentation of the results. We hope this helps to prepare the reader about the structure and ordering of the presentation]

5) Figure 10—are you serious? Again, this is not needed in-text since you've lost your hardcore readers by this point.

* Ok. Thanks and relegated to appendix.

6) I suggest the authors tackle the issue of endogeneity tests in this piece in supplementary materials. More detail needs to be included regarding diagnostics—their dismissal of them again suggests this paper was initially prepared for a different format. Given the authors' desire to provide guidelines for practitioners, discussion of other available measures—again in the context of IV assumptions—would further solidify the contribution of this ms. Current treatments across a range of estimators are inconsistent with regard to presenting Hansen's J, Sargan, Bassman, Cragg-Donald; all are measures readily available in existing software packages. For instance, for the Hausman test to be reliable, the model must be appropriately specified. The Hausman endogeneity test performs poorly in the presence of weak instruments (in addition to Staiger and Stock 1997 see Hahn and Hausman 2003 in the American Economic Review), a point the author does not address. Also, the finite

properties of these tests have been examined in Kirby and Bollen 2009 and Guggenheim 2005.

* [Thanks. This point bears resemblance to the point bout validity vs reliability, so please see also our discussion above. Following R3’s suggestion, we are now discussing endogeneity tests in a more structured way. The Hausman test is really not relevant here but we are now citing all relevant literature. Importantly, we think we already discuss the limitations of the tests the reviewer refers to in section 4. There, we also explain why are not primarily focused on these tests in this paper. As we wanted to underline both in the introduction and in the, this paper does not make any direct contribution about validity; rather it makes a contribution about how one can possibly increase first stage strength conclusion (see e.g. new footnote 3 on page 3 and first and third paragraph of conclusion of the revised MS). Although we believe these points are very useful, there is not enough space to talk about everything here. We wanted to focus on the issues that relate to the contribution of the paper. We believe that a more general piece, such as by Sovey and Green or by Bollen would be better suited to also discuss best practices regarding validity.]

7) Advise the readers to 'eyeball check' the graphical fit? What would the reader anticipate it to look like based on your description to this point?

* [We thank the reviewer for this point. We now make this suggestion more explicitly in section 5, where we discuss the empirical examples of the paper (pages 30, 35).]
* Test for non-linearity. Given the KRLS results the reader needs to verify that the nonparametric estimation captures the salient features of the relationship between x and z.

8) The third empirical application was best executed of the three in the context of why a political scientist might be interested in an instrumental variable estimation approach—there is suspicion of reverse causality and a need to construct a model that will enable the researcher to construct and empirical model that will allow estimation and assessment. This is the hook if you will for the readership of the journal and for most uses of these techniques. Reshaping the ms to take into account an explicit subfield will improve the manuscript immensely.

Bollen, K. 2012. "Instrumental Variables in Sociology and the Social Sciences." Annual Review of Sociology 38:22.1-22.36.

* [Thank you for this comment. Focussing on one subfield will reduce readership especially since the audience of AJPS is a general one. The examples also serve to indicate that IV estimation has spread from economics through almost all subfields of political science.]

**Reviewer 2:**

The authors attempt to improve on 2sls in the context of the generic model

(1) Y = α + βx + ,

in which cov(x,) is not equal to zero thereby invalidating least squares as an estimator of β. (Rather than invoke the question at the many points where it is relevant, I’ll note that throughout the discussion, Angrist and Pischke notwithstanding, none of the estimators discussed in this paper are unbiased estimators of anything. The discussion is about consistency, not unbiasedness, and the distinction matters. Throughout this paper, the discussion floats back and forth between the two notions as if they were synonymous.)

* Thanks for this comments we now very carefully distinguish between unbiasedness and consistency, and fully separate finite sample from asymptotic properties.

Also, a small point that would matter if it were not an obvious oversight. In equation (2), the constant term α is not the same as α in equation (1).

* Thanks of course. This oversight is now rectified

Finally, throughout the paper, the authors refer to the “OLS model” and “2SLS model.” OLS and 2SLS are estimators, not models. (See, e.g, the diatribe in Wooldridge’s most recent undergraduate text.)

* Yes thank you, we have changed the wording throughout

This specification thus far invokes an instrumental variable approach. Thus, the authors suggest the always present other equation,

(2) X = δ + γz + u.

This is treated as a structural equation in the model. I would note, in the contemporary treatments (see, e.g., those of Wooldridge or Imbens), this implicit other equation in the system is taken only to be the the projection of x on z, not a structural equation. That is all that is needed for the principle of instrumental variable estimation to proceed, e.g., the Wald estimator in equation (4).

* This is correct and so noted. We now distinguish more carefully between the econometric correct notation and the practical application which usually is a linear equation in the first stage of the 2SLS.

The argument in this paper is ostensibly about efficiency. But, much of the discussion on pages 5 and 6 is about “bias.” (See the note above.) Equation (4) does not demonstrate that the IV estimator is biased, and the comment about the correlation of Z and X does not suggest a bias in equation (5) either. There does seem to be some commentary about the numerator term in (5). The comment vaguely suggests that if the correlation of z and  is nonzero, that the IV estimator will be inconsistent. The elements of (5) are population quantities, not finite sample elements. It is not clear how the population correlation (zero) becomes nonzero in a finite sample.

* Thank you we have carefully revised all the notation in the paper and made clear where we refer to population characteristics and where to finite sample quantities. We also more carefully distinguish between discussion of bias and efficiency. We now formally show that a nonparametric first stage generates an estimator that maximizes efficiency in the first stage.

The end of page 6 suggests, broadly, that instrumental variables should be correlated with the variables they are instruments for. There is a peculiar statement, “Sometimes our prior intuition or theory about the effect of about the effect of z on x holds only locally; higher values of z may be associated with higher values of X … This seems like a nonsequitur. First, when does this happen? Second, why is it relevant to this development?

* Thank you for this comment. We have rewritten large parts of this section to make clear where efficiency gains come from and why a nonparametric first stage is better suited in an IV estimation.

The development thus far treats (2) as if it were specifically a structural equation – earlier statements suggest something of interest about δ. As noted, for the theory of the IV to work, (2) needs only to be viewed as the linear projection in the joint distribution of X and Z. However, section 2 turns to alternative “representations” of this relationship, namely x = f(z) + v, nonparametrically. Three candidates are suggested, kernel smoothing, local linear regression and kernel regularized least squares. With respect to the third of this, the text states “the KRLS estimator allows the data to represent the relationship between outcome and covariates.” (There is only one covariate here.) I cannot figure out what this statement means. If there is a “relationship,” what is it. Is it different with different data?

* Thank you for this comment. We are now discussing nonparametric approaches more clearly and in more detail. For KRLS we now use both econometric representation and also MC DGPs with more than one covariate in the main outcome equation.

Section 3 turns to “Statistical Properties.” This is very difficult to unpack. It must be the case that X and u in (2) are uncorrelated. The authors seem to mix uncorrelated in the population with uncorrelated in the sample.

* Thanks, see comment above. We now carefully distinguish between population and finite sample properties

The residuals in the least squares regression of x on z in (2) will be orthogonal to z even if z and u are correlated in the population.

* That is true

The quote from MHE mixes this notion. To be specific, even if E[x|z] is some nonlinear function of z does not mean that d + cz cannot be a valid instrumental variable.

* This is true and we are not claiming otherwise, but we clarify the discussion in this section.

Moreover, even if “the true conditional expectation function is actually a probit or logit,” it will still be ambiguous what is meant by a “clean residual.”

* Yes the text has not been very precise and we rectified this.

The discussion then goes on to suggest that none of this is relevant to the nonparametric approaches because they don’t make functional assumptions. They are just functions of the data. Surely if this is the case, why is it not possible to use some arbitrary polynomial, or even a probit function while just denying that the model is a probit model, and Φ(z) is just an approximation. After all, the normal CDF can be approximated very well using a polynomial in z. This argument is not persuasive. In any event, this section should be the crux of the matter.

* Elias – I am not sure here. I don’t know whether this is actually correct. Any idea?

It is argued that the proposed estimators based on a nonparametric first step are preferred to 2SLS because

(1) The uncertainty in the assumed functional form in (2) is carried over into the second stage.

(2) The method proposed here minimizes the number of instruments at 1, and more instruments increases the finite sample bias of 2SLS. I would note, this is a moot point in this study, as there is one regressor in the entire model and one instrumental variable used throughout.

This leaves what must be an open question here. There really is no convincing story told for WHY the techniques proposed here should “outperform” 2SLS.

* Thank you. And sorry that we have been unclear, and this is our fault. These two points are not the main reason for using a nonparametric first stage. We have rewritten this section to make clear that the whole point about a nonparametric first stage is that it maximizes efficiency w/r to OLS in the first stage without violating the orthogonality assumption between u and z.

A note on methods, “In a typical 2SLS model [sic], this is done by including the R2 from the first stage in the variance in the variance of the second stage OLS estimator.” This is not true. It is a pedagogical device used in Wooldridge’s undergrad text. The correct variance for 2SLS is computed by using OLS at the second stage, then computing the residuals (or their squares) using the original X, not the predicted X. Moreover, it has nothing to do with normality as suggested later in the paragraph.

* Okay! We rectified this.

A criterion for comparing the estimators using a sort of bootstrapping method is suggested on page 13. This needs to be explained more clearly, then it should be justified as a measure of the performance of the estimator.

* Sorry that we were unclear. We are not using bootstrap to compare performance at all but to include uncertainty from the first stage into the second stage of the 2 stage IV model.

Is the “Mean error variance of the second stage” the variance of the replicates of the second stage estimators of β1 around the true variance? The second term is the variance of the coefficients across all K samples. Isn’t this the same as the first term? Obviously not. It’s not clear what it is, or why this second term is relevant to a comparison the sampling variation of the second stage estimator, which is what is at issue here.

* We have dropped the bootstrap discussion since it doesn’t seem to be relevant at all for recovering accurate SEs in the nonparametric estimation procedure.

Section 4 presents the Monte Carlo results. The discussion begins with the claim that “Both 2SLS and the nonparametric 2-stage estimator should outperform simple OLS in case x is indeed endogenous and z is a valid instrument for x.” This is simply not true.

* Okay point taken. We change this wording so it is accurate

It depends on the underlying correlations, of course, but it will always be the case that the variance of OLS around its mean is smaller than that of 2SLS around its mean. Whether the MSE is smaller is open, of course. Several dozen figures in this analysis then go on to show that by the criterion in (12), the nonparametric estimators are vastly to be preferred to OLS or 2SLS. (I’d like to see a reconciliation of (12) with the suggestion on page 13.)

* Thank you for this point. We are now discussing bias and efficiency separately
* Unbiasedness vs. efficiency

This entire analysis is done in the context of as trivial a model as it is possible to devise. I’d like to see some suggestion as to whether the authors believe there is any generality here.

* We have expanded the MCs accordingly, but there are space constraints and we have relegated much of the results to the appendix

For example, it is stated on page 26, “when the true relationship is nonlinear or even non-monotone, nonparametric versions of the two stage IV model (sic) with the first stage lowess or kernel always outperform the 2SLSestimator even in case the instrument is very weak and the bandwidth is very small. In these cases, lowess smoothing slightly outperforms kernel regression (11). Is this intended to be a general result? If so, it is pretty powerful and deserves more comment.

* Yes this seems to be general, and carries through all the additional MCs that we have now undertaken. We are also including some more theoretical statistical discussion in the earlier sections to make this clear.

Once again, however, I’d be interested at least in some speculation as to why this should be true. A similar discussion applies to the bottom of page 29, where it is speculated that the standard errors produced by one of these estimators are too large. Is this general? How so? Why?

* Thanks again. We are explaining and discussing this now in more detail.

From pages 31-39, three applications from the literature are explored. In the first one, the new estimator is compared to the received results based on some kind of R2 measure for the main equation. I don’t believe this would have been anticipated by any of the preceding analysis. (There is a bit of pedantry in the text in this paper. On page 32, why is the continuous IV “quasi-continuous” and how does quasi-continuity relate to the exercise done here? Likewise, in the second application (about Reagan’s popularity) it is not clear what is being compared here. It seems to be suggested that the results of the original study were nonsensical: “Accordingly, it comes as no surprise that the final 2SLS estimates we get are completely uninformative.” Is this the original results or the newly generated results in this study. Again, however, none of these has anything to do with the efficiency of the estimator.

* We take this point. We are more closely relating the discussion of the examples to the theoretical and MCs results. We apologize that we have not been clear enough but all of this has to do with the efficiency of the estimator. Since it captures the nonlinear relationship in the first stage better it is more efficient and thus produces a much smaller partial R2.