RESPONSES TO REVIEWER

MAJOR COMMENTS:

1. **Comment:** ***"My main concern with the paper is that it does not fit the format suggested for the replication section as described in Economics E-Journal's Replication Guidelines. More specifically, the authors do not make an effort to actually reproduce and report RY's results. … the authors should include RY's results without the correction (in a way directly comparable to their results) in the paper."***

   Response: Following the reviewer's comment, the revised version now includes a "replication" section that replicates RY's key results without attempting any correction. See Section 3 on pages 10f.

2. **Comment:** ***"It would be nice if the authors provided some descriptive statistics of how much the autocorrelation in the constructed x changes due to their correction and how that better reflects real-world datasets (currently the autocorrelation in x is not reported in the paper)."***

   Response: Table 5 of the revised manuscript demonstrates the extent of the excessive autocorrelation introduced by RY's experimental design. This is accompanied by a discussion in the text that elaborates on the nature of RY's mistake and our correction (see pages 11f.)

3. **Comment:** ***"[The] question arises whether RY's recommendations still hold for real-world datasets with high autocorrelation in x, i.e. whether in cases with high autocorrelation in x heteroscedasticity in the errors should be taken into account for estimator choice."***

   Response: We did not attempt to explore the limits of RY's recommendations from artificially induced autocorrelation, but chose instead to focus on the kinds of datasets one is likely to encounter in actual empirical datasets. As indicated from Table 1, our analysis includes a wide variety of diverse datasets. As indicated by Table 5, some of these datasets are characterized by a high degree of autocorrelation.

MINOR COMMENTS AND QUESTIONS:

4. **Comment:** ***"Related to major comment 1: Why do the values in Table 2 for the RY datasets do not correspond exactly to the values in RY's Table 2? Shouldn't sigmas and rhos be the same if the same datasets were used to produce them?"***

   Response: There are two reasons for this. First, the simulated panel datasets are created using a (pseudo) random number generator. This introduces sampling error in the simulated datasets even when the population data generating process (DGP) is the same. The second reason is that the values in Table 2 (now Table 6) are estimated conditional

on the explanatory variables, and our paper uses the "corrected" explanatory variables that do not incorporate excessive autocorrelation. Even so, the results are still very close.

5. **Comment:** ***"The description of how the sigmas and rhos are constructed based on the real-world datasets is not entirely clear. Do the authors use the first N observations and then all contiguous T combinations for these observations or do they use all possible combinations of N cross-sectional units and with all contiguous T combinations?"***

Response: Please see the discussion on the top of page 4 of the revised manuscript.

6. **Comment:** ***"While I like the authors idea to try to find a way to find a uniquely recommended estimator for both estimation and inference, I think the bootstrap exercise they present at the end of the paper does not really achieve this goal: The authors show that Estimator 7 (the recommended estimator for estimation for the reported cases of T/N combination) with bootstrapped standard errors outperforms Estimator 8 (the recommended estimator for inference). But can one be sure that Estimator 8 with bootstrapped standard errors would not perform even better than Estimator 7 with bootstrapped standard errors in terms of accuracy? They should at least note that based on this exercise they cannot uniquely recommend Estimator 7 but as this estimator's accuracy is very high with bootstrapped standard errors it is probably ok to use it."***

Response: It was never our intention to suggest that the Parks with bootstrapping had better inference performance than the PCSE estimator with bootstrapping. We rewrote this section in a manner that we hope makes this clear. Page 19 states, "While only an example, this exercise suggests that when $N \leq T$, a single-estimator approach that uses the Parks estimator with bootstrapping can be superior to the two-estimator approach that relies on the Parks estimator for coefficient estimates and the PCSE estimator for hypothesis testing. This is a topic for future research." We hope this statement avoids any misunderstanding.

7. **Comment:** ***"The comment on IV in the conclusion seems a little out of place: I am not sure the authors should sell it as a contribution of their paper that they can make recommendations based on observed data characteristics while this is not possible in other situations. It would be a contribution if it had not been possible before to give recommendations based on observed characteristics for the problem that they pose themselves (i.e. cross-sectional dependence in panel data settings) – but it seems weird to sell it as a contribution when compared to entirely different problems."***

Response: We agree. The reference to IV estimation has been removed.

8. **Comment:** ***"Figures 1-8 (particularly 3 to 8) are hard to read in black and white print. It may help to introduce different shapes for the different lines that the reader is supposed to look at (and combine this with color coding if necessary)."***

Response: We experimented with different shapes for different lines but decided to stick with the reliance on color coding. Using both different shapes and different colors made the graphs too "busy" and difficult to read.

9. **Comment:** ***"The paper should be spell-checked and checked for duplicate words and punctuation errors."***

Response: We have spell-checked the document and checked for duplicate words and punctuation errors.