Overall referee report response (original points in *italics*, abbreviated for brevity):

Response to Referee #1:

1. *My first comment is that the first two items [in the proposed replication preanalysis plan] seem to be the same...*

    (a) The three items of flowtime, budget, and accuracy are distinct items. Flowtime refers to calendar days needed to complete the project, whereas budget includes all accounting costs plus working hours. For example, after a project begins, a day spent not working on the project counts towards flowtime, but does not affect the budget's working hours.

2. *My second comment is motivated by the comment that "the amount of flowtime and budget that you could invest in a replication could grow uncontrollably." The wording strikes me as unusual...*

    (a) I have reworded this paragraph to be as follows, which emphasizes stopping rules per the editor's suggestion: "Prespecification of the amount of flowtime and the budget for a replication would provide you, as a replicator, with explicit stopping rules. These stopping rules would be useful because proving that a paper is not replicable under any circumstance could require a very large investment, as you would have to show that all possible permutations of your decisions in your replication would lead to the conclusion that the paper is not replicable."

3. *Finally, I wish to stretch Andrew's recipe a bit, perhaps beyond his intent, and comment on the "precommitment" literature...*

    (a) As I mentioned in my previous response, it was indeed my intent to tie this paper to the "precommitment" literature. In the specific context of replications, I believe that preanalysis plans can both mitigate incentives for replicators to try to debunk original research articles and also provide replicators with a results-free defense against criticism from authors. More generally, the benefits of preanalysis plans for original research also apply to replications.

Response to Referee #2:

I added a link to the call for papers to this draft in the front matter. For your specific points:

1. *I fail to see how the statements made about replication attempts do not apply more generally to essentially all scholarly research projects.*

   (a) In response to your point and also the concerns of referee #1 and the editor, I have reworded the first statement's paragraph to emphasize stopping rules, which now reads as: "Prespecification of the amount of flowtime and the budget for a replication would provide you, as a replicator, with explicit stopping rules. These stopping rules would be useful because proving that a paper is not replicable under any circumstance could require a very large investment, as you would have to show that all possible permutations of your decisions in your replication would lead to the conclusion that the paper is not replicable." I have also included an additional argument to the introduction as to why preanalysis plans are particularly beneficial for replications, namely that preanalysis plans mitigate incentives replicators might have for debunking original research papers.

2. *Several arbitrary and/or unnecessary criteria are included in the selection method. For example:*

   - *a. The statement about not selecting from the authors own previous replication work is a pure redundancy from the second (more general) criteria, which stated that previously replicated papers were to be selected. By definition, if the author had previously replicated a study then someone had previously replicated it.*

     – I have clarified in footnote 9 that these statements are not redundant.

- *b. To exclude work by those with a connection to the current place of employment and those with a personal correspondence history seems completely arbitrary.... Again, I am not challenging the idea that this 'comfortable' method is a fine way to proceed with picking a study to replicate, but my point is that any 'comfortable' and/or 'clearly non-scientific' selection method does not merit independent publication as a stand-alone academic contribution.*
  - Although the editor requested that I remove the non-mandatory selection criteria from the paper, for maximum transparency I have kept them in the current draft.

3. *Several arbitrary and/or unnecessary criteria are included in the definition of "success". For example:*

   a. Point 4 is quite vague. The author would wait a "prespecified amount of time" (later quantifying as a 'few' weeks) and would engage in a "prespecified number of attempts" (never quantifying).

   (a) I have amended the preanalysis plan to quantify these points.

   b. Similarly, point 5 mentions a "flowtime of around two months" to do the replication... Would the author's own life/circumstances that played out during the time in question not be accounted for in a reasonable way?

   (a) I have added a contingency clause to the proposed preanalysis plan (step #7) to accommodate this point.

   c. Most importantly, point 9 indicates that "If the data that I downloaded was obviously flawed, then I would give up and work on another research paper." The mind (ok, MY mind) reels upon reading this. I trust the authors of the original Haurin and Rosenthal paper because I have no reason at all not to. I trust the author of this paper for the same reason. Setting that aside, it seems that in the case of a purely doctored *research endeavor, this step in the process would actually cause the research conducting a replication to give up when they should not...*

   (a) In response to concerns from the editor and referees about the example preanalysis plan, I have reworded my example plan to be for the context of replicating Haurin and Rosenthal for verification for the archival record. This step no longer exists in the preanalysis plan, though I have also added a clarifying statement at the beginning of section 3 that says that the proposed plan presumes the authors did not undertake a doctored research endeavor.

4. *While I understand the researcher, and in turn the journal to which the paper is submitted to, carries a direct interest in the field of Economics. However, Economics has direct connections to other disciplines including Finance, Accounting, Marketing, Political Science, Geography, Sociology, Urban Planning, and many others. [A simple review of the 750+ journals indexed by Econlit substantiates this point.] In every way, the paper was written as if other fields do not exist.*

   (a) I argue for using preanalysis plans that specify flowtime, budget, and intended results for "success" in replications. I was not thinking directly of economics per say. But as economics is the field that I am most familiar with, the references fall within economics, and the special issue's requirement mandated selection of an economics article to apply the proposed replication procedure to.

Response to Referee #3

1. *When the different contexts are described I miss an explanation of how "a verification of the original paper for the archival record" is meant. Why should one keep such a record of others' papers? For the journal editor? Why are online appendices irrelevant?*

   (a) What I was referring to by the "archival record" was the portion of the document that an institution (e.g., journal, library, or preservation society) is going to preserve. In the revision, I have clarified this context to indicate that, if a replicator would be trying to verify material for the archival record, that the replicator would be interested in all estimates that were designated for preservation. Appendices could be designed for preservation, but it is also possible that they would not be designated for preservation.

2. *The problem with pre-analysis plans in empirical economic research when data is available is that there are many fast and easy ways to engage in data mining that would not leave any traces and could easily have been done before the pre-analysis plan was published.*

   (a) Yes, your point is true. I would hope that researchers would be ethical enough to abstain from data mining prior to writing a preanalysis plan, if the researchers are intent on following such a preanalysis plan.

   (b) Let me touch on a related point: I think that exploratory research, i.e., research undertaken without preanalysis plans is valuable. Exploratory research is necessary in certain contexts, particularly when the researcher is so unfamiliar with the data (or is unfamiliar with data generating process) that the researcher is unable to craft a preanalysis plan. I also think that, on net, preanalysis plans are underused in economics.

3. *While it is specifically mentioned that pre-analysis plans should help to ground our estimates in statistical theory I don't see any theoretical reasoning for the replication.*

   (a) My comment that preanalysis plans can help ground our estimates in statistical theory is applicable to both new research papers and replications, not just replications per say. I have added a citation in footnote 4 of the revision that discusses pretesting.

4. *The author mentions an extension but does not specify in what way he plans to extend*

the original study.

   (a) In response to your concern and the concern of the editor (in the editor's letter, points # 6, 7, and 8), I have substantially revised the replication plan. The revised plan now outlines a replication procedure for verifying Haurin and Rosenthal (2007) for the archival record instead of for the purposes of writing an extension.

5. *One google scholar citation is a questionable measurement of influence.*

   (a) There are undoubtedly better measures for "influential", but a citation seemed like a reasonable litmus test. Papers submitted to the special issue had to preselect an "*an influential economics article that has not previously been replicated*" and gain approval for the article by the special issue's editor prior to writing our submissions. Therefore, I have kept my originally selected article in the revision.

6. *Checking the ReplicationWiki is not sufficient to determine whether a study has already been replicated.*

   (a) It's difficult (or even impossible) to prove a negative, so a registered replication on the ReplicationWiki seemed like a reasonable check to see whether this paper had been replicated. Looking over the citing articles in Google Scholar as of December 4th, 2017 I did not see any article titles that obviously suggested that a replication had been conducted (and the only paper with "replication" in the title that also cites Haurin and Rosenthal 2007 is this paper).

7. *For the replication plan I would suggest to mention which software the author plans to use.*

   (a) I have added to the replication plan that I would try to match the software version-operating system combination that Haurin and Rosenthal used.

8. *...the amount of time for the original authors to respond to inquiries for their material and the prespecified number of attempts that the author would try to contact the original authors should be specified.*

   (a) I have quantified the number of attempts and provided a more detailed description on the procedure through which I would contact authors in the revision.

Response to Referee #4

1. *"(iv) a discussion of how to interpret the results of the replication (e.g., how does one know when the replication study "replicates" the original study).": In my view, this is the only weak part. One of the author's three criteria for a replication plan is to "set (...) the set of estimates and the degree of precision that will define a "successful" replication." (lines 3 and 4 of the article). In his proposed replication plan, step 10 on page 8 reads: "I would be "successful" if I was able to replicate the Figures (...) to a reasonable degree of accuracy." This is quite vague and, in my view, does not meet the author's own criterion.*

   (a) In the revision I have rewritten the replication plan to be more precise about the degree of accuracy needed for a "successful" replication (steps 10 and 11 in the revised plan), including a metric for replicating the figures from Haurin and Rosenthal (2007).

Response to Annette Brown

Generally, I agree with you that the definitions of 'success' and 'failure' are somewhat gray. I use the term 'success' to match the special issue's language and I use the terms 'failure' or 'unsuccessful' because they are opposites of 'success'. For your specific comments:

1. *I think [the three contexts for replication sucess in the introduction] would be more useful if you extended them to talk a bit about what replication procedures you would employ in the three situations and not just what counts as successful.*

   (a) The purpose of the three examples was to emphasize that the context for a replication matters for ’success’ and ’failure’. The procedure employed for each context would also be different, but differences in procedures is not the focus of these examples, so I have left them as they were in the previous draft.

2. *Along these lines, you do not say anything about whether and where replication preanalysis plans should be posted publicly.*

   (a) In the case of original research articles, I’m not sure whether a public posting or a private posting of a preanalysis plan would be better. Public postings of original research preanalysis plans essentially give other researchers the ability to free ride on the careful preanalysis plan, but the benefits are, of course, more transparency and the ability for the public to criticize or comment on the public plan. In the case of replications, however, I think that because the cost of public posting is reduced (i.e., less need for others to free ride) public posting is probably a good idea. I have added a sentence to the draft in footnote 2 on this point.

3. *One remedy that we introduced for this [replication taking longer than the plan anticipated] is requiring of 3ie-funded replication studies, or recommending for others, that replication researchers always begin with a push button replication.*

   (a) I agree that needing to be able to do a “push button replication” could be an important step in a preanalysis plan. I don’t think that a replicator should be mandated to add such a step to a plan; it would be up to the replicator to decide.

4. *I think you could usefully add a lot more discussion here [in the replication plan], including:*

   (a) *How did you come up with the estimates that you have here?*

      i. The flowtime and hours estimates are based on my own experience. I have added a footnote 11 clarifying this point.

(b) *What is the breakdown of the timeline and working hours by some key milestones in the replication work (in your case, maybe by figure in the original paper)?*

   i. I'm not convinced that having a milestone breakdown of flowtime and hours in a preanalysis plan is a good idea figure-by-figure. Presumably, when replicating multiple figures, the marginal cost is decreasing per figure, but it's difficult to anticipate by how much the marginal cost would decrease and the benefit of prespecification figure-by-figure is questionable, at least to me. I agree that a researcher could specify contingencies whereby the replication would cease before the maximum allotted flowtime and budget are met (i.e., if certain milestones are met or not met). I don't know whether figure-by-figure would be the appropriate metric.

(c) *What happens if the timeline or working hours are fully expended and you haven't been able to fully replicate? I personally would not argue that a replication should be considered failed if it cannot be completed in the planned time (either duration or working hours), but I also think there needs to be a stopping point so that a replication researcher does not spin her wheels.*

   i. I disagree on whether to consider the replication a failure if it cannot be completed in the planned time. The replication should be considered a failure, *conditional* on the prespecified amount of flowtime and budget if the replication were undertaken using a preanalysis plan that specified such variables. In this case of a replication failure, the paper could still be replicable under *some* circumstance (e.g., more flowtime or budget), but it would be a failure in the circumstances specified in the preanalysis plan.

5. *I agree that sometimes you do need to "give up", but there should still be an output in this case.*

   (a) I agree that there should be some output in the case of failed replication, a point also commented on by the editor. I have clarified in step #11 that I would report any output that I produced (regardless of whether I would be successful or unsuccessful).

6. *I appreciate the pre-specification of the research process that you present in section 3 of the paper, but I am more interested in the analysis plan, about which there are very few details... Will you try to code using their estimation methods as described in the publication? Will you look to other sources to learn more about their methods (i.e. is there a working paper with additional information)?*

   (a) In my view the replication should be sufficient based on the information that is in the publication and the citations therein. In the case of Haurin and Rosenthal (2007), their estimation procedure cites Maddala (1983), so the details of the estimation (and related data transformations) should be self-contained in those two papers. I have added details to the paper in step #1 of the revised replication plan on these points.

7. *Finally, if you propose to use the label successful, as you suggest in step 10, then I think you need to be more precise about what you consider to be a "reasonable degree of accuracy".*

   (a) I have modified the plan to be more clear on the level of accuracy for success in steps 10 and 11.