Response to Referee

I thank the referee for the many thoughtful comments. These have made me think more deeply about some of the issues, which would improve the paper.

General Comment

I will start off with a general comment, which may clarify some of the issues, before responding to the specific comments of the referee. My thinking regarding replications has evolved as a result of working on this project, as a result of the ideas presented in the other papers, and as a result of the referee's comments. I have come to see that if replications are going to become more mainstream, they need to serve an identifiably useful function, and to do that, the core part of a replication would need to be standardized with a rather narrow focus, which should include the two parts that you refer to as a "push button replication" and a "pure replication" (Woods and Vasquez 2017). As part of this, there should be an examination of the data, an examination of the code, and comments on anything that is an important influence on the results.

These different parts would be answering the following questions:

- Can the author-provided data and code yield the published results? Yes/No Comment.
- Are there any issues with the data that, if altered, would significantly alter the results? Yes/No Comment.
- Are there issues with the code that, if altered, would significantly alter the results? Yes/No Comment.
- Are there any other issues that, if changed, would significantly alter the results? Yes/No Comment.

Once this foundation is laid, then it is possible to pursue the nobler goal of replication "as part of the process for translating research findings into evidence for policy" (Brown et al 2014, 220). Replicators may want to avoid the use of the terms "errors" or "mistakes," but I would think that research that is found with serious errors or mistakes in these first two steps, the push button replication and pure replication, can be dismissed as not credible and, therefore, there is no need to go further. I agree that there are some gray areas, but at some threshold, some research can or should be dismissed.

The major concern that I am dealing with in the paper is that if replications become more common, and, as many point out, that replications are more likely to be published if the results refute the original, then there should be some rules as to what it means to refute the original to reduce the incentive to find some basis on which to say that the original is wrong. I am suggesting that significant errors, ones that affect the conclusions of the original paper, in these first two steps can be seen in more or less black and white terms. Differences that arise after that point are where I would suggest the gray areas arise, the purpose of the exercise being that replications can provide an incentive to increase the quality of published research by adding clarity not confusion.

It is also worth noting that while 3ie has the incentive to produce replications that help to identify research that can be translated into policy, the demands of academic publishing produce an incentive structure that may produce replications with a less worthy goals, that being to refute findings whether warranted or not in order to get published. My thoughts are to try to reduce that problem.

This paper outlines how to approach replication research from a theoretical perspective and then applies that theory to an empirical case. While the author provides some general details for his approach, I think additional specifics would greatly strengthen the contribution of the paper to the transparency literature. After spending 5+ ...[more]

... years running 3ie's replication programme, I found myself referencing a few of our studies on this topic. I limited myself to 3 personal references. Please feel free to cite other references if you find those more useful.

Comments:
Page 2: When describing "what is true" I struggle with the concept of one absolute truth. Isn't there space for grey?

This is meant to be an idealized scenario. The point is that the replications should add clarity.

Page 2: I would argue in almost all situations no researchers are infallible. I would suggest there is value in replication research that does not necessarily determine the "truth" but provides an alternative researcher's perspective on the data and analysis.

I agree that there is value in alternative perspectives, but I think that replications should add clarity rather than confusion. If more outlets become available to publish replication studies, and the replication studies that are more likely to get published are ones that refute some aspect of the original, then this may create an incentive that promotes more confusion about the credibility of the original rather than more clarity about its value. Given that possibility, it should be absolutely clear to readers where the replication of the original ends and the alternative perspective begins.

Page 2-3: I would like to see a deeper discussion of how the author defines reliable results, i.e. constructive criticism as briefly described in the current draft.

This is a fair criticism. My feeling is that the threshold for "reliable" depends on the step of the replication. At the push button stage, the results should be virtually identical with identifiable reasons for any differences. I don't think that differences at this stage would necessarily mean that the research is found unreliable, but the expectation should be that results will be virtually identical. In the second step, which includes an examination of both the data and the code, the results should support the conclusions of the original.

Page 3: How does a replication researcher (or a consumer of replication research) define "duplicate" results. Do the results need to match exactly, down to each decimal point? What if software or user written code differences (for example) result in "minor" differences? Would something like the push button replication protocol that we're currently working on be helpful in this situation? http://www.3ieimpact.org/media/filer_public/2016/07/13/replication-protocol-pbr.pdf

This is addressed in the previous comment.

Page 3: Where the author mentions reaching out to the original authors, is there a suggested framework for such a conversation? What incentive do the original authors have to participate in this process?

My somewhat limited experience may not be the norm, but I have found authors to be generally willing to share data and code when it is not otherwise available, and when I have let them know what I am planning to do. I recognize that there are situations where authors may be unwilling to share data and have very little incentive to do so.

Page 3: I would caution against thinking reliable results need to be exactly identical. Please provide additional details here.

While I agree with the view of 3ie that the overall purpose of replication is not to uncover errors and mistakes, finding serious errors and mistakes is a possible outcome. If this is the outcome, the question then becomes whether there is something wrong with the original or whether there is something wrong with what the replicator did. I would think that the consumer of the replication would have more confidence in the results of the replication if the replicator can come up with the same results or an explanation for any differences. By starting with the original data, code, and an open dialogue with the original author, the replicator reduces the errors he/she may introduce and increases the credibility of the replication if it finds problems with the original. Any other basis for finding a problem with the original seems unfair to the original author given the bias to publish replications that refute the original.

Page 3: I would also like to think that positive replication findings can independently verify the original findings and give policymakers more confidence in the original results. I'm not seeing a space for this type of situation in the author's description of quantifying replication findings.

I don't think the original author will have any problem with positive findings nor will the consumer of the replication be necessarily confused by the positive findings. Clearly, positive replication findings strengthen the case for the original study.

Page 3: The author's description of the different steps he proposes for a replication study reminds me of our "Quality Evidence for Policymaking: I'll believe it when I see the Replication" paper. I agree that there should be a clear delineation between reproducing the original results (both using the original data/code and trying to recode the methodology using the original data and the publication) from robustness checks or additional analysis. I would the author describe these processes in a bit more detail. With all things replication related, the devil really is in the details.

The previous comments, I think, deal with this.

Page 3: I found the researcher's motivation for the study selection really lacking. The researcher could have picked a number of different papers to develop a replication plan, why did he choose this one?

I will totally agree with you on this one! I was attracted to this paper because it seemed to be making a big claim about a big issue. Corruption undermines growth and development. It seemed interesting if the innovation of social media was providing a social benefit in reducing corruption. The paper also pulled together data from a number of different sources, which served one purpose in being able to comment on whether the data is assembled correctly. Errors are sometimes introduced when data is

just linked together using what may seem to be a unique identifier, which is why I am suggesting that critical examination of the data should be part of the process.

Page 4: I didn't understand the "except for religion variables" mentioned around the Religion Data Archive. Aren't all of the variables being used religion-related variables?

This sentence may be poorly worded. The Religion Data Archive variables are religious variables and they are for the year 2005. The other variables used in the regression are all for the year 2012.

Page 4: I found the conversation around the "correct (negative) sign" counterproductive to replication research. I would argue the whole idea of replication research is to attempt to verify the original results, but I wouldn't call those original results necessarily correct.

This part is a summary of the original paper. The word "correct" is indicating that it was sign that supported the original author's conclusion. Perhaps "expected" would be a better word to use here.

Page 4: I didn't understand the end of the second full paragraph, when the researcher discusses "methods used that differ from the choices made in papers they reference." Aren't we most interested in differences made by the replication researchers that differ from the original methods/analysis?

I think one of the issues that needs to be dealt with is that published research often presents findings in the way that best supports the author's thesis. In developing a replication plan, it is fair to ask if the original author made choices that differ from ones made in previous research that is being cited, and whether these choices affected the results in some way. This sentence is referring to the development of a replication plan stage rather than the stage where the replicator may be introducing different approaches.

Page 4: Before assessing arguments, I would claim that the starting point is testing if an independent researcher can run the code on the data and generally reproduce the original results. If not, than discussing the paper's original argument seems unnecessary to me.

This part was not developed very well. My thought here was much like what you describe as "we would start by drafting a replication plan" (Wood and Vasquez 2017, 2). For this, I would look at what the original author is basing the conclusions on. This would be a starting point to understand the original paper and for developing the plan as to what to focus on in the replication.

Page 4: How would the author suggest replication researchers go about checking assumptions? Which assumptions should be checked in this proposed replication paper?

The assumptions that I was thinking of are the ones that involve the data chosen and the methods used for analyzing the data where there are choices that the author seems to be dismissing. The candidate paper mentions other sources of corruption data and other approaches that have been used in previous research to look at the relationship between internet use and corruption. Your papers bring more to the table on this issue.

Page 4: While I agree that it is difficult to recode an original publication from scratch, and I can attest to

that from my past research, I thought the paper dismissed this concept a bit too quickly. I believe it is the job of the original authors to describe transformations and corrections in their publication, working papers, or supporting materials. While it might be very difficult to obtain identical results, I would hope a reasonable researcher could generally reproduce very similar findings.

My concern is with the basis for rejecting the original study, while you have the nobler goal of identifying useful research. I think that yours is the more important of the two, but if replications are going to become more common, then there should be a clear dividing point between a basis to reject findings and disagreements over the importance of findings.

Page 5: I would like to see some organization to the questions at the top of the page. Our "Which tests not witch hunts: a diagnostic approach for conducting replication research" in this issue provides a systematic approach for these types of replication questions. You might find this paper or others that have attempted to categorize approaches to replication research helpful.

I do find your paper to be helpful in developing some of the ideas touched upon here.

Page 5: I didn't understand the danger being described in the first full paragraph on this page.

I see that I am coming at this issue from a different perspective than you are. Ideally, yours is the better perspective, but I also think that there is also a role for replications that simply validate original results. A.N. Brown et al (2014, 224) in their discussion of pure replications point out that this "is important for validating the original results and is also the necessary first step for further replication tasks." What I am suggesting is that if a replicator is presenting results that suggest a different conclusion from the original author, then it should be clear what stage of the replication process this is based on. It seems to me that the further the replication strays from the original code and data, the easier it is to dismiss the replication if it comes to a conclusion that refutes the original in some way. I am not suggesting that identifying differences at later stages isn't useful, but I am suggesting that given the incentive a replicator may have to get something published in an environment where it is more likely for a negative result to be published than a positive result, the less credible a negative result may be.

Page 5: I kept expecting to see policymaking recommendations highlighted somewhere on this page but it never appeared. Is there a reason why?

Policymaking recommendations should be the ultimate result, but that went beyond the scope of this paper, which was to lay out an approach to replication.

Page 5: I would think any chances to the methods should also ideally be pre-specified, no?

Yes. That would be ideal.

Page 5: Who judges "failure" to replicate or "assembled incorrectly"? How should we think about quantifying these types of comments? I'm assuming original authors and replication researchers would have very different thoughts on this regard.

My thought here is with the noncontroversial stages of the replication, which you refer to as the push button replication and the pure replication. At these stages, there are results that could be obtained

from the replication that would convince most people that the results of the original are not valid. The most convincing would be results that don't support the major finding of the original. The original author may not agree, and this is where I am of the opinion that if the replication is to add clarity rather than confusion, the disagreement should be on issues that are pretty much black and white.

Page 5: I found concept of "render[ing] results meaningless" difficult to fathom. Similar with the idea of "results that do not undermine the study." This language needs to be clarified.

The point that I am trying to make is that the finding of failure to replicate should be restricted to a relatively narrow basis, which this paragraph is attempting to outline. If the main conclusion of the original study is based on the significance of a coefficient of a particular variable, then if that doesn't hold up under the scrutiny discussed, then there could be the finding that that the replication failed.

Page 5: I believe there is more grey space in replication than exactly reproduces and failed to replicate. Is there a reason the author has taken such a straight-line approach?

There is much gray area in all of this. I think that the incentives that 3ie face make it more credible to dive into these gray areas, but it should be recognized that these may be different from the incentives that other replicators may face. That being the case, the underlying question is how to produce reliable replications in an environment with a bias towards publishing replications that refute rather than confirm the findings of the original. My suggestion is to standardize the core portion of the replication, as outlined previously, and based only on this portion can one say that the replication has failed, if that is the finding. Anything done after that can deal with the gray areas.

Page 6: I would like to see a more detailed conclusion. I find value in verifying original results and think more confirmatory replications would actually help change the dynamics around replication research. When discussing replication studies "casting doubt" on the original results, I would like to see answers to the "where", "how", and "who judges" types of questions. What if the original authors disagree on if the replication study casts doubts on the original study? I can provide many examples from 3ie's replication paper series if the author would like to explore this issue at a greater depth.

The ultimate judge would be the consumer of the replication study. I would be interested in information about any disagreements between original authors and replication authors and then any policymaking decisions that resulted. That could be very interesting.

Thank you again for all the effort you put into reviewing my paper. This has been a very useful experience for me.

Sources:

Brown, A.N., D.B. Cameron, and B.D.K. Wood (2014). Quality evidence for policymaking: I'll believe it when I see the replication. Journal of Development Effectiveness, 6 (3): 215-235.

Wood, B.D.K and Maria Vasquez (2017). Microplots and food security: encouraging replication studies of policy relevant research. Economics Discussion Papers, No 2017-71, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-71