Comments to be posted publicly and anonymously

Referee Report for **MS 2355** – 11/3/2017

"Microplots and food security: Encouraging replication studies of policy relevant research" by Benjamin Wood, International Initiative for Impact Evaluation and Maria Vasquez, independent consultant

In this review, I was asked to focus on two questions: (i) Is the contribution of the paper potentially significant? (ii) Is the analysis correct? Given that this is a somewhat unique paper, providing a replication plan, but no actual analysis, rather than focus on the "correctness of the analysis" I focus on aspects of the paper which need clarification or suggestions to strengthen the paper. The majority of my comments have to do with specificity—as a replication plan should be sufficiently specific as to understand with precision the analysis to take place. In terms of the first question (contribution of the paper), I assess the potential for contribution to be significant, particularly as it is part of a special issue on replication (an area where more work is needed in Economics). Overall, I find the paper to provide a useful discussion of how replication could further the policy-relevancy and enrich the results of existing published research on an important development topic.

Specific comments to be addressed by the authors are as follows:

- **The first section** of the paper "A general discussion of principles…" is quite dense and includes a lot of information—it also starts quite abruptly with no introduction to the rise and/or relevance of replication to the field of Economics. The narrative may be easily understood by someone familiar with the replication process in general, or the 3ie replication process, specifically. However, to someone unfamiliar, the narrative may need a bit more context and/or examples throughout. For example, these could include: a few sentences about 3ie's replication work/policies, explanation of how pre-decision rules around classifications (major, minor) are made, examples of what types of measurement or theory of change robustness checks could "look like." It is understood that this is a short article, and thus the article is not meant to comprehensively address these issues, however a bit more context would be helpful for the non-expert reader.
- **Section two, page 3**: The narrative around the importance of women's land ownership/rights could be strengthened here—as this was a critical component of the program. This could be done easily by citing or expanding the narrative a bit using this new DP: Meinzen-Dick et al. 2017: http://www.ifpri.org/publication/womens-land-rights-pathway-poverty-reduction-framework-and-review-available-evidence
- **"Push button replication"**: Given the number of indicators that Santos et al. report in Table 3 (18 in my count)—it is not clear why the authors specifically choose the four indicators as "key results" for the PRB (listed bottom of page 4). Since the ToC described in Santos et al. is quite general, these four indicators do not seem particularly more relevant than the others. Further, given that there are variations in significance across the indicators/models presented in Santos et al. Table 3, the selection of what constitutes a "key result" seems like an important decision, which needs more explanation.

- **“Pure replication”**: With regards to the propensity-score weighted regression technique, it would be useful to mention what additional information the authors were expecting to find in Santos et al. which was omitted – both in the article and Appendix material (“We will also pay special attention to examine the inverse propensity score score-weighted regression technique used in the original analysis to estimate the casual effect, as it is not discussed at length in the paper.”) As this is a fairly common technique—does it need extensive explanation? What would be additional critical information for readers to know regarding the application of the method? As an aside, should this be “inverse propensity score-weighted regression” (do you need the extra “score”?)
- **Section 3.1: Research assumptions: attrition**: In Santos et al. it states that “We were unable to interview 338 of these women and had to replace their households . . . “ this seems clear to me that if no woman was present, the household was not administered the survey. However, if the authors think this is ambiguous, they are fair in suggesting inclusion of households without females available to be interviewed as a robustness check. In this same section, the authors should spell out more clearly what exactly their further differential attrition checks would be (beyond what is provided in Santos et al. including appendix material).
- **Section 3.2: Data transformations**: The robustness check suggested by the authors for household hunger cannot be directly tested, as the question asked was only one question referring to both the scenarios “did not have food and/or money to purchase food” – however since this is not made sufficiently clear in the article, this is a fair robustness check to propose at this stage.
- **Section 3.3: Estimation methods**: It would be useful for the authors to further explain the “annual analysis” – how would this differ econometrically (and in spirit) from the analysis conducted by Santos et al. on heterogeneity by “Moved to NGNB plot” (Table 4). In addition, does the suggested district-level analysis involve separate analyses by district? Are the authors not concerned about the small sample sizes that will results from this analysis—and the potential loss of power? Or perhaps I am misinterpreting what “district level analyses” means here. Some clarification would be useful.
- **Section 4.1: Heterogeneous impacts**: It would be useful to note here how wealth will be (potentially) measured.