Dear Garret,

Thank you for this helpful feedback. We found your suggestions useful and we plan to incorporate them into the revised version of our paper we will later submit.

Please find our responses to your comments below. We would be happy to further clarify anything if you have additional questions.

Sincerely,
Ben & Maria

Pg 2: I assume "We would then assess the reproducibility of the paper by recoding the original results, the pure replication stage" means recoding from scratch, or from the earliest point possible? Authors should clarify exactly what that means—going back to primary data sources, or just re-writing code, etc.

Yes, we mean recoding from scratch or the earliest point possible the entire paper. In the pure replication stage we would use the data set provided by the authors to conduct the data analysis without using their original code. Instead, we will code the analysis based on the paper's description of the analysis, any additional description of the analysis in the working paper, and any descriptions of the data analysis in possibly available readme files associated with the data.

Pg 3: A more complete justification of why this particular paper was selected instead of others on a similar intervention would help. The program evaluated is obviously policy relevant, but as the authors indicate, similar programs have been implemented elsewhere. Were they evaluated, and are there papers on them? If no other papers exist, state that.
A brief summary of the analysis conducted in the original paper would be very useful for the reader

Thank you for this feedback. We will incorporate a deeper summary of the original paper and more motivation for our paper selection in the revised submission.

Santos et.al. (2014) conduct an analysis of the the effects of distributing micro plots on intermediate and longer run food security outcomes using baseline and endline data from 1035 households: 671 of them received the program and 364 were controls. Given problems of attrition between the baseline and endline and balance issues across the treatment and control group, the authors use an inverse propensity score-weighted regression model to estimate the effect of the intervention. The analysis is conducted at the household and the plot level (for tenure security outcomes). They also test if the size of the plot or including the name of a woman in the land title influences the outcomes of interest.

In regards to why we selected this paper, empiric research of the effects of land transfer programs is rare. As far as we know, this paper is unique in the sense that it analyze the effects of distributing micro plots of land on food security. Given the scarcity of land availability in many

countries, we believe assessing the effectiveness of distributing micro plots is more policy relevant than many of the other similar interventions on this topic. Although a couple of other papers evaluate the effects of land transfer programs, the one's we're aware of assess the effects focusing on income not nutrition (for example Keswell and Carter (2014) and Benjamin, Brandt, McCaig, Le Hoa (2017)).

Pg 4: Is "iv) Household used seedlings, seeds, or grafted stems in last year" a good measure of use of improved inputs? Please define these terms, as at least to this non-farmer, it's curious how seeds are considered an improved input since they seem like the required input in essentially all farming. Also, some justification of why the four chosen indicators are the ones chosen from the roughly twenty in the original paper's Table 3 might be helpful. The authors refer to the theory of change, but a more specific discussion of why the other sixteen are less important could help.

In our minds, each group of indicators (tenure security, use of credit for agriculture production, investments in agricultural production, women's participation in decision making) provide valuable information about the effects of micro plots on nutrition. We selected the outcome within each of the four groups that we considered most relevant to the theory of change. In the case of "tenure security" we focused on the women's belief around having access to the plot in the following years. For "women's participation in decision making" we thought that at least the women should have a say on how to use a share of the land to be able to grow nutritious food. Regarding the "use of credit for agriculture production" we determine that the household should report that they got a loan from the bank in order to invest in the new plot.

Finally, for "investments in agriculture production", we agree that use of seeds is a prerequisite for most agricultural activities. We did not a have a strong preference for the improve inputs indicators. We selected the used seed outcome because we considered that the household should at least invest in seeds, considering that the majority of the households were previously landless. Also, we chose this measure measure based on the original researchers' logic. They note in their footnote 14 that this measure quantifies the likelihood of households "to undertake new plantings and/or annual crops, rather than only caring for already existing trees and perennials." (p. 871) All that being said, we're open to alternative argues if researchers have different suggestions on this point.

Pg 5: I agree the propensity score is a potentially important avenue of researcher degrees of freedom, both for the original researcher and the replicator. Please provide more detail on how this would be investigated.

Based on the the numerous propensity score techniques, we agree that propensity score matching remains an area ripe for replication and reanalysis. We originally envisioned testing the robustness of the inverse propensity score-weighted regression to alternative matching techniques. But given the lack of information on the original approach, we are simply proposing here that we would provide detailed information on the original approach. We plan to follow King

and Zeng (2006) to test the common support assumption and Austin (2017) to examine the balance between the treatment and control households.

Pg 6: Investigating the adult/child definition sounds good. Please provide a citation or definition of adult equivalency unit.

Adult equivalency in the literature is oftentimes used within labor economics to assess relative welfare within households of different compositions (see Nelson (1993) for an overview). In this instance we plan to use adult equivalency units closer to the way Frongillo and Nanama (2006) measure food insecurity. They cite Swindale and Ohri-Vachaspati (1999) as a general resource for approaching adult equivalents in this context.

Pg 6: "We plan to clarify with the original authors…" This is the first mention of contacting the original authors. I might be useful to provide general information in the intro section about a plan for interaction with the original authors. To get the original data and code would one have to contact the original authors, or is the data publicly available? (I couldn't find it in a quick search.) These are important general questions in the area of replication research.

The original data are not publicly available. For this replication we would request i) original data and code, ii) information about the software use for the analysis, and iii) any replication instruction to the authors by e-mail. During the PBR and pure replication stages of our analysis we would consult with the authors if we find that the original data or code is not complete (or major discrepancies appear) to confirm that we have the right code and data set. We would also share the final report with the authors for their optional comment.

Pg 7: The discussion of land deciles seems unclear. I assume decile is referring to some percentile of some distribution, but it is unclear what distribution that is. Or is decile a unit of land measurement? I have never heard of that term, but that seems to be the way you describe it: "households that received 5 or less deciles of land and 10 or less deciles of land."

Thank you for the detailed reading of our paper. We actually need to correct a misstatement here. The word "decimal" should be included instead of "decile". A decimal is a unit of land measurement (1 decimal =1 /100 acre).

Bibliography
Austin PC. 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*; 46:399–424.

Benjamin, D., Brandt, L., McCaig, and Le Hoa, N, 2017. Program participation in a targeted land distribution program and household outcomes: evidence from Vietnam. *Review of Economics of the Household.* https://doi.org/10.1007/s11150-017-9390-0

Frongillo, E and Nanama, D, 2006. Development and Validation of an Experience-Based Measure of Household Food Insecurity within and across Seasons in Northern Burkina Faso. The Journal of Nutrition; 136: 1409S-1419S.

King, G and Zeng, L, 2006. The Dangers of Extreme Counterfactuals. *Political Analysis*; 14: 131-159.

Keswell, M and Carter, M, 2014. Poverty and land redistribution, In Journal of Development Economics; 110, 250-261. https://doi.org/10.1016/j.jdeveco.2013.10.003.

Nelson, J, 1993. Household Equivalence Scales: Theory versus Policy? Journal of Labor Economics; 11: 471-493.

Swindale A and Ohri-Vachaspati P, 1999. Measuring household food consumption: a technical guide. Washington: Food and Nutrition Technical Assistance Project, Academy for Educational Development.