Summary

> The author motivates his replication of Mu and van de Walle (2011) by highlighting the importance of the original study and by referring to a call by Brown, Cameron and Wood (2014) in the "Journal of Development Effectiveness" to validate empirical findings. First, the author describes the motivation, data and method of the original study which is followed by a pure replication (re-estimation the original model specification and documentation where differences occur). Next, the author 1) modifies the estimation of the logit model (adding pre-treatment outcomes and dropping insignificant variables), 2) changes the choice of the bandwidth of the kernel estimator, 3) analysis two additional outcome variables. The pure replication reveals several deviations from the original results in the descriptive statistics: four out of 24 above 10%. These differences are reflected in the results of the estimated treatment effects, though the conclusions of the original article remains unchanged. Similar, modification 1) and 2) do not change the original conclusions. The two additional outcome variables (credit and migration) added as extension 3), are not affected by the treatment.

General comments:
To my understanding, the central challenge in evaluation studies is securing the appropriate counterfactual. One aspect of this challenge is the correct application of the chosen statistical technique (in this case PSM). The authors focus on this aspect and leave aside other aspects (e.g., explicit testing of plausible alternative explanations for the measure effect, definition of treatments, covariates or outcomes). Since the original article is not very detailed on alternative choices in the PSM, this can be justified as a focus of the replication.
The authors analyze how a modification in the variable choice for the logit model and the bandwidth for the kernel matching affects results. This choice seems ad-hoc, as there are many more modifications the researcher could analyze and it is not justified why these two are chosen. There are several guidelines which document the choices the applied researcher faces (e.g., (Caliendo and Kopeinig, 2008) (Imbens, 2015))). From this point of view, the reader is let alone in deciding if the affect of the most crucial assumptions were chosen to be analyzed.

More specifically, it would be relevant to learn more about the matching quality (see e.g (Caliendo and Kopeinig, 2008)) beyond what has been been done in the original article. A sensitivity analysis (see also Caliendo and Kopeinig, 2008) would have provided the reader with a systematic understanding of how failing the unconfoundedness assumption would influence the conclusions. Other aspects include, how the common support is determine or why kernel matching and PSM are chosen as methods.
One way how to organize a systematic replication could be similar to our recent replication of an article base on PSM (Lampach and Morawetz, 2016) where we follow the guidelines by (Caliendo and Kopeinig, 2008). But there certainly are alternative good approaches for a systematic replication of PSM.

Specifically I was asked to answer the following two questions:
Question 1: Is the replication done to a high standard of professional competency?

As far as re-estimating the original results is concerned, I consider the replication having a high standard of professional competency. The author seems to have a good knowledge of PSM. The description of how this this done, though, is too detailed to be interesting for the reader in my point of view.
But the interesting aspect of a replication is not so much if the identical analysis is possible, but if the results hold under alternative specifications which are key to fulfilling the assumptions for unbiased results. In this respect, the article lacks a systematic approach as described in my general comments above.
Question 2: Do the robustness/extensions add value to understanding the original study?
To my understanding a robustness check would require to 1) specify the crucial assumptions and then 2) check the robustness with respect to these assumptions. This gives the reader certainty about how credible the results are. As the manuscript lacks such a systematic presentation of the assumptions, the reader cannot be sure the results or the original study are credible even after reading the replication study.
The extension (adding two additional outcomes) are of limited additional value as it is not explained why these two outcomes should be affected by additional roads in the first place (other than "These outcomes are important for livelihood and non-farm diversification of rural households, and can provide policy-relevant findings") and how to interpret it that there are no significant treatment effects.

From my point of view, the current manuscript lacks substantial elements of a replication study with substantial value for the readers. It would be possible to modify the manuscript in such a way. But this would require a substantial re-structuring of the article and a lot of additional analysis. It would thus rather be a new submission than a revision.

Minor Comments:
It would be good to make code and data available to the reviewers and readers. This would add to transparency of the research.

At several places in the manuscript there are references to the appendix. In the version of the manuscript I got, there is no appendix.

p.2 line 5 from bottom: probably "... they also find heterogeneity in the impact of ..." instead of "... they also find that heterogeneity in the impact of ..."

p.3 line 8: probably "Mu and Van de Walle (2011) provide important findings on the role of rural roads in non-farm employment and market development." instead of "Mu and Van de Walle (2011) provide important findings on the important role of rural road in non-farm employment and market development."

From the end of page 2 to page 4, four reasons are given why the Mu and Van de Walle (2011) article is important. But is this the only reason why this study was chosen for replication? I would expect that the author also found some aspects not convincing which made him want to replicate?

p. 7: The details of the .do files the author got from Mu and van de Walle are not very relevant for the reader, I think. I agree that it is less than optimal if descriptive statistics don't match, but it is not so important to dedicate it a separate table to it, I think.

At places the manuscript reads more like a story: "I found a variable of the predicted propensity core in the data sent by..." or p. 10 "I checked the data carefully, but cannot find the reason...".

p.16, line 5: One could argue, that PSM also depends on the function form, because the propensity scores are calculated with a parametric logit model.

p.17 (ii) I guess it should be "Several control variables are statistically insignificant in Mu and van de Walle (2011)" instead of "Several control variables are statistically significant in Mu and van de Walle (2011)"

References

Caliendo, M., Kopeinig, S., 2008. Some Practical Guidance for the Implementation of Propensity Score Matching. J. Econ. Surv. 22, 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Imbens, G.W., 2015. Matching Methods in Practice: Three Examples. J. Hum. Resour. 50, 373–419. doi:10.3368/jhr.50.2.373

Lampach, N., Morawetz, U.B., 2016. Credibility of propensity score matching estimates. An example from Fair Trade certification of coffee producers. Appl. Econ. 48, 4227–4237. doi:10.1080/00036846.2016.1153795