Original authors' feedback on "Impacts of Rural Road on Household Welfare in Vietnam: evidence from a Replication Study"

This paper reports on a replication of our 2011 study "Rural roads and local market development in Vietnam" by Ren Mu and Dominique van de Walle, Journal of Development Studies 47(5) 2011.

Replication studies are dependent on the availability of a certain amount of documentation on what was done in the original study. But this too applies to one's ability to judge a replication study. It is difficult for the reader to judge the replication given the paucity of detail on what was done exactly, including how variables were constructed.

The replication focuses primarily on the (very) small differences found in the attempt to replicate. The key discrepancies appear to arise from the following two factors.

The first important difference is that the replication uses 198, not 200 observations for the PSM logit regression. This means that the propensity scores and the common support are slightly different, so that different estimates result. The differences are fairly small and it is no surprise that they are there. The author notes that he drops the 2 communes due to missing values in some explanatory variables. He doesn't say which variables have missing values. But, based on the data we have, two variables – the share of crop land and share of perennial crop land – have missing values for two observations in 1997. Assuming that these are attributes that are relatively sticky over time and given that they are not of interest as outcomes, we replaced these with the values for the same communes in 1999 and were able to run the regression with 200 observations. This seems the obvious thing to do. We suspect that if the authors of the replication study had done so then the replication would have been more exact.

The second key difference stems from the author's different definition of a few outcome variables. We were able to find all the necessary variables and reproduce the same numbers as in the published article for adult illiterate, credit availability, men's barber, women's hair dressing, primary school completion and secondary school enrolment rate (the variables that differ). We have no idea why the replication study could not get the same numbers. We had men's barber and women's hair dressing coded separately while the replication has them as one variable. The primary school completion in the replication study is way too high and has a puzzling decreasing time trend. Ours started with about 31% in 1997 and increased to 37% in 2003, which seems a more sensible trend, given the time and context. We suspect there are errors in the replication study, but beyond these observations it is hard to say what they might be.

In sum, the replication is using a different sample (based on a different PSM logit regression) and occasionally differently defined variables. It is no wonder that the results are not exactly the same. It is difficult for us to say much more about the replication.

As well as presenting its own estimated variable means and estimated impacts, the paper reports the difference between its replicated estimates and those given in the original paper by noting the difference between them as 0%, <10% or >10% difference. Although most estimates have 0 or less than 10% difference between them, this way of presenting the results tends to exaggerate the differences.

For example, in one case the author reproduces the original table 1 which gives mean baseline characteristics for communes classified by median household per capita consumption, and gives his own version (also as Table 1). In most cases, the means are the same. In many of the <10% difference cases, the differences are miniscule and look like they could be due to rounding off errors. For example, for Market availability, this paper reports 0.31 and 0.66 compared the original paper's 0.32 and 0.63. For bicycle repair shop, it is 0.54 and 0.88 versus 0.53 and 0.88. In only four cases, the means are very different and are undoubtedly defined differently. As noted above, in the case of women and men's hairdressing services, the variable is aggregated while the original paper reported them separately. For the other three, different definitions have clearly been used by the original study and the replication.

The results part note in various places (e.g. page 12, page 18, page 19) that "most of the impact estimates replicated in this study have the same sign" as in the original study. That too gives the wrong impression. Not only are they of the same sign, they are often the same or extremely similar. The paper exaggerates small differences.

In the end, the paper concludes that the differences are due to differences in the construction of the variables and not due to methodological issues. What is remarkable is how little the qualitative conclusions alter and this is surely notable. The replication finds no faults with any of the do files or the methods used. The only problem is with some of the data cleaning documentation. The replication study might also comment on the degree to which the original paper provides details on definitions and what it has done. It is obviously not perfect but compared to most published papers it is quite detailed. We would like to see the paper focus not only on the very small differences but the incredible similarity of the results.

The replication tests sensitivity of the original results to changing the bandwidth in the kernel matching and to adding different covariates to the logit model used to compute the propensity scores. None of these tweaks changes the results. However, we are not sure that what is done in the second change makes much sense. First, the baseline value of each outcome variable is added singly to the logit model and propensity scores (PS) computed. Thus a different logit model and PS are estimated for each outcome. Common support presumably alters at times too. This seems a very strange thing to do. First, the original logit already contains baseline proxies for most outcome variables. Second, the PS is meant to estimate the probability of each commune getting the road project. This will clearly not vary by outcome variable. Finally a balancing test showed that baseline outcomes are similar after matching. The paper has not followed standard practice in these respects.

The paper also argues that the logit should be pruned of all covariates that have lower than 10% statistical significance citing a paper that argues that "inclusion of irrelevant variables can increase the standard error of the estimates." But the fact that a covariate is not statistically significant does not imply that it is irrelevant. Clearly the characteristics of poor Vietnamese communes in 1997 are likely to have been highly correlated and as a result insignificant in the logit model. This does not mean that they should be excluded. One would need to be very careful in deciding what attributes were or were not relevant.